Chapter 5

How to decrease the risk of unreliable and invalid evaluations?

Abstract

Implementation of effective teacher evaluation procedures is a global challenge in which lowering the chances that teachers receive inaccurate evaluations is a pertinent goal. This study investigates the minimum number of observations required to guarantee that teachers receive feedback with modest reliability ($\boldsymbol{E}\rho^2 \geq .70$) and that any summative decisions about their professional career have high reliability ($\boldsymbol{E}\rho^2 \geq .90$). A sample of 198 classroom observations by 62 colleagues of 69 teachers working at eight schools reveals that reliable feedback requires at least 4 lesson visits by four different observers. Also results indicate that if only using classroom observation it is almost impossible to guarantee a reliability level sufficient for the use for summative decisions. The findings mirror those reported with other observation instruments. This study accordingly offers directions for how schools can implement classroom observation procedures cost-effectively.

Modified version of the article: Van der Lans, R. M., Van de Grift, W., J., C., M., Van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation, 50,* 88-95. doi: 10.1016/j.stueduc.2016.08.001

5.1 Introduction

The development and implementation of effective teacher evaluation is a global challenge, as various international policy documents and reports reveal (e.g., DfEE, 2012; Mourshed, Chijioke, & Barber, 2010; NCTQ, 2013). In all these policy documents teacher evaluation has a dual purpose: (1) identification and selection of ineffective teachers and (2) offering advice for improvement of teachers' teaching (Marzano, 2012). The global attention signals that many countries currently are interested in how to obtain more reliable information to support their summative decisions and formative feedback. That is, there is an interest in preventing wrong decisions about teacher selection and preventing the provision of wrong feedback about how to improve teaching effectiveness, since wrong decisions and feedback will harm individual teachers, and definitely not improve student learning outcomes.

Of these two purposes of teacher evaluation, the decisions about teacher selection currently receive most attention (e.g., Firestone, 2014; Winter & Cowen, 2014). Evidently, there is much at stake for individual teachers, who have worked hard to earn accreditation and succeed in classrooms. This gives researchers and policymakers the moral obligation to carefully consider the reliability of their decisions. Clearly, evaluations might be wrong and select teachers for dismissal which will prove to be effective. Also, evaluations might be wrong by not selecting teachers for dismissal which will prove ineffective. Currently, it is attempted to avoid wrongly removing effective teachers, but this automatically leads to a situation in which many ineffective teachers are wrongly retained (e.g., Winters & Cowen, 2013).

The provision of formative feedback has at first sight less severe personal consequences. Nevertheless, also feedback should be based on a representative picture of the teacher's true teaching skill. In general, educational policies rely on classroom observations specifically to target teachers who appear ineffective in some way and to provide them feedback (e.g., NCTQ, 2013). If these teachers show no improvement in their follow-ups, the policies suggest they should be selected for dismissal. Given these personal consequences, teachers deserve reliable feedback, such that it offers them a true opportunity to improve.

This study examines the reliability of classroom observation. Classroom observation is currently the most widely adopted teacher evaluation method (Strong, 2011). However, only few studies report on the reliability of these observation methods (e.g., Hill et al., 2012; Kane et al., 2012) and none of these studies relate reliability criteria to the two

different purposes of teacher evaluation. This study seeks to determine if classroom observations can achieve a reasonable level of reliability to support both formative feedback and summative decisions, and if so, how many observations by how many separate observers are required.

5.2 Background

5.2.1 Evaluation reliability and purpose

An examination of validity and reliability should be related to the purpose for which the instruments will be used (Kane, 2006). In teacher evaluation, instruments generally are used for two different purposes. Therefore, different reliability criteria should apply to investigate whether instruments reliably support summative and formative evaluation decisions. However, studies examining classroom observation instruments rarely relate reliability criteria to the intended use of the instrument. For example, Hill et al. (2012) examine how much the reliability increases if evaluations incorporate multiple raters and lessons and seek "to achieve acceptable reliability" (p. 60), without clarifying what an acceptable level of reliability would be and whether that level might change if other evaluation purposes would apply. Similarly, Kane et al.'s (2012) influential report for the Measures of Effective Teaching (MET) project notes that:

> *"Not all decisions require high levels of reliability. Measures could be used many different ways: promotion decisions, retention decisions, compensation decisions, or low-stakes feedback intended to support improvement. Different uses necessitate different evidentiary standards and different levels of reliability (there is no uniform standard that applies to any envisioned use)."* (p. 13)

That is, though Kane et al. (2012) recognize that different evaluation purposes require different reliability criteria, they do not mention any specific criteria. In subsequent work for the MET project, Ho and Kane (2013) cite the reliability criterion $\boldsymbol{E}\rho^2 = .65$, without specifying the evaluation purpose for which this criterion would be appropriate. Because these studies do not set clear reliability criteria for different evaluation purposes, it appears that the reliability of classroom observations currently is determined by educational policies and the school principals' perceptions of what it takes to get a "reliable observation" for a given purpose.

To tie evaluation purposes to different reliability criteria, we adopt the criteria for both modest and high reliability formulated by Nunnally (1978). Therefore, we argue that modest reliability of $\boldsymbol{E}\rho^2 \geq .70$ suffices for formative feedback and other instances in which the stakes are relatively low. We suggest instead that a high reliability level of $\boldsymbol{E}\rho^2 \geq .90$ is the minimum criterion to use for summative decisions and instances in which "a great deal hinges on the exact score made by a person on a test" (Nunnally, 1978, p. 245).

5.2.2 Reliability of one-time lesson visits

Using multiple lesson visits is not standard practice in teacher evaluation, with some notable exceptions, such as the teacher advancement program (TAP) (Darling-Hammond, Amrein-Beardsley, Heartel, & Rothstein, 2012; Toch & Rothman, 2008). Yet it is common knowledge that one-time observations may be substantially biased by a bad moment or difficult class (e.g., Muijs, 2006; Shavelson & Dempsey-Atwood, 1976). In empirical studies of the reliability of a single lesson visit by a single observer, across different classroom observation instruments, the findings are fairly consistent. Ho and Kane (2013) report reliability coefficients between .27 and .45, depending on the type of observer (teacher peer or administrator); Kane et al. (2012) examine five classroom observation instruments and report coefficients of .37 or less. In Hill et al.'s (2012) study, the reliability coefficients for three different subscales of the Mathematical Quality of Instruction (MQI) hover between .37 and .46. That is, the reliability of single classroom observations is low and generally less than .50. Previous works suggest that at least three to four lesson visits are required to achieve even modest reliability ($\boldsymbol{E}\rho^2 \geq .70$) (Hill, et al. 2012; Ho & Kane, 2013; Kane, et al. 2012). Note that we use the notation $\boldsymbol{E}\rho^2$ to refer to the reliability coefficient. This notation is taken from Brennan (2001). The ρ^2 is the usual notation of reliability in classical test theory. The $\boldsymbol{E}$ signifies that the reported coefficient reflects the expected reliability. It is the reliability we would expect if the evaluation procedure is exactly repeated.

Beside low reliability, the validity of one time classroom visits has also been criticized on other grounds. One is that the person visiting also is the person judging and hence observation scores cannot be anonymous (Scriven, 1981). This makes the appointed evaluator most vulnerable to criticism (Popham, 1987; French-Lazovik, 1981) which in turn provides an incentive to give lenient scores (Centra, 1975; Weisberg, Sexton, Mulhern, & Keeling, 2009). Both Centra and Weisberg stated that an evaluation procedure which

evaluates over 95% of the teachers as performing sufficient lacks validity. These studies show the necessity to clearly distinguish between those who observe and those who decide.

5.2.3 Potential evaluation procedures

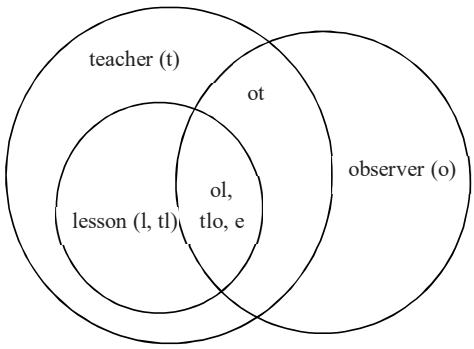
With the view that reliability is paramount to teacher evaluation and that single-lesson visits have unacceptably low levels of reliability, we discuss three evaluation procedures that might enhance the reliability of classroom observations, compare their pros and cons, and speculate whether their durable implementation in schools is realistic. The successful implementation of any evaluation procedure requires that it be cost effective and manageable for schools (Peterson, 2000). Ideally, an evaluation procedure would entail minimal organizational complexity but still provide sufficient guarantees that the resulting evaluations are reliable and fair. Furthermore, any implementation is restricted by the reality of the school organization. We consider three potential procedures: crossed, nested, and bias-confounded.

Crossed procedure. This complex evaluation procedure requires a group of observers to visit all lessons together. An example of the crossed procedure appears in Figure 5.1. At the left side of Figure 5.1 the evaluation procedure is visualized. Check boxes reflect that the observer visited the lesson.

Figure 5.1

A schematic representation of the crossed evaluation procedure (left) and the resulting variance decomposition (right)

	Observer		
	A	B	C
Teacher A			
lesson 1	✓	✓	✓
lesson 2	✓	✓	✓
lesson 3	✓	✓	✓
Teacher B			
lesson 4	✓	✓	✓
lesson 5	✓	✓	✓
lesson 6	✓	✓	✓
Teacher *n*			



At the right side, a Venn diagram representation is used to visualize the same procedure. In a Venn diagram each circle is a facet; areas where two circles overlap

illustrates an interaction between two facets. The crossed procedure offers the most complete information, because it separates information about true differences across teachers (t) from any bias due to differences across lessons (l), bias due to observers (o), and bias due to their interaction (observer × teacher). In our notation, the "e" refers to "error." Furthermore, commas identify confounding facets. Confounds signal that variation is attributable to two or more facets such that the variation has no single interpretation. Hence the facet "lo, tlo, e" in Figure 5.1 reflects that this part of the variation in scores may be explained by lesson × observer interactions, by teacher × lesson × observer interactions, and by measurement error. As such this facet has no substantive interpretation.

This crossed evaluation procedure has been applied in previous studies of the reliability of classroom observations (Hill et al., 2012; Ho & Kane, 2013). It offers benefits, in that the crossed design offers information about the reliability of the evaluation, as well as details about the extent to which any particular bias affects reliability. If reliability is too low, the procedure reveals what to do: (1) add another observer, (2) prevent some particular observer from visiting some particular teacher, or (3) visit an additional lesson.

Despite its comprehensiveness, this evaluation procedure is unworkable in practice for most schools. In the hypothetical scenario where a school employs 50 teachers and requests three lesson visits with each teacher, it would demand 150 group visits by the same group of observers. The number of work hours also depends on the size of the group, but in this hypothetical case, if the group includes three observers, it would mean 450 hours of lesson observation. Most schools lack the financial resources to hire external observers, so the observation group likely consists of peer colleagues, team manager(s), or school principal(s). Each of these actors would have to perform 150 classroom observations, in addition to their existing obligations, and schedule these observations together. It is implausible that such procedures can be implemented successfully in schools, despite that this would be better from a psychometric point of view. In addition, likewise the one-time lesson visit procedure, also an appointed group of 'expert' observers will be vulnerable to criticism (French-Lazovik, 1981; Peterson, & Chenoweth, 1992). Because in the crossed procedure all teachers are evaluated by the same (small) group of observers and the observers will be more acquainted with some subjects, or befriend with some colleagues, it is likely that some of the teachers under evaluation will not feel that they are treated equally. Note also that in a research setting the strength of the crossed procedure is that it can take such observer-teacher interactions into account, but in the school practice there is

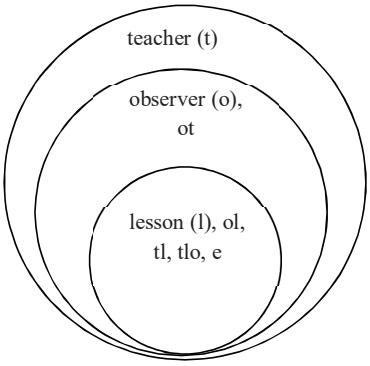
no knowledge about such statistical models and currently it is unlikely that schools can take adequate actions to avoid tensions between colleagues when implementing the crossed procedure.

Nested procedure. As a more flexible approach (Figure 5.2), the nested procedure requires one group of observers to visit multiple lessons of one teacher together. The difference with the crossed procedure is that other teachers may be visited by other groups (see Figure 5.2). This flexibility comes with a price though. The procedure cannot reveal the extent to which reliability decreases due to observer × teacher interactions. Rather, the variance due to observer × teacher (ot) interactions sums with the variance due to observers (o), resulting in an "o, ot" facet that confounds two interpretations. That is, the variance in this facet might reflect differences among observers, or it could reflect differences in observer × teacher interactions.

Figure 5.2

A schematic representation of the nested evaluation procedure (left) and the resulting variance decomposition (right)

	Observer					
	A	B	C	D	E	F
Teacher A						
lesson 1	✓	✓	✓			
lesson 2	✓	✓	✓			
lesson 3	✓	✓	✓			
Teacher B						
lesson 4				✓	✓	✓
lesson 5				✓	✓	✓
lesson 6				✓	✓	✓
Teacher *n*						



None of the research referred to in this study has used the nested procedure. It offers benefits in that it is more flexible with regard to who can perform the classroom observations in comparison to the crossed procedure. This flexibility is important since it provides the room to carefully select peer-observers for each teacher (French-Lazovik, 1981). Furthermore, it still provides some information about what to do if reliability is too low: add another observer or visit an additional lesson. However, the nested procedure is not any more efficient than the crossed procedure. Its implementation in our hypothetical, modest sized school would require different groups of observers to visit 150 lessons

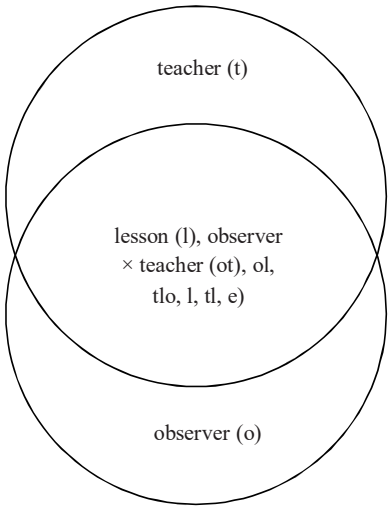
together, so if we again assume the groups include three peers, it still demands 450 hours of observation. Also, despite that now different groups may perform the classroom observations, schools still have to schedule group visits. They need to find groups willing to visit lessons together.

The bias-confounded crossed procedure. A yet lesser complex procedure involves the bias-confounded crossed procedure. In this procedure, teachers are grouped and teachers within the same group visit each other's lessons (Figure 5.3). The term "bias-confounded" signals that in this procedure all interaction facets are summed. The main difference with the nested and crossed procedures, is that a bias-confounded procedure allows for individual lesson visits. Thereby, the bias-confounded crossed procedure is much more efficient and requires only one-third of the lesson observations (i.e. 150 hours of observation) compared to the previously described crossed and nested procedures.

Figure 5.3

A schematic representation of the bias-confounded crossed procedure (left) and the resulting variance decomposition (right).

	Observer						
	A	B	C	D	E	F	G
Teacher A							
lesson 1		✓					
lesson 2			✓				
lesson 3				✓			
Teacher B							
lesson 4	✓						
lesson 5			✓				
lesson 6				✓			
…							
Teacher E							
lesson 7						✓	
lesson 8							✓
lesson *n*							
Teacher *n*							



The increased efficiency comes with a cost. In comparison to the crossed procedure, if using this procedure, it is not possible to estimate the size of bias due to variations across lessons (facet l). The facet "l" is summed with the facet observer × teacher interaction (ot) and all its interactions "ol", "tl" and "tlo" and further confounds with measurement error. In

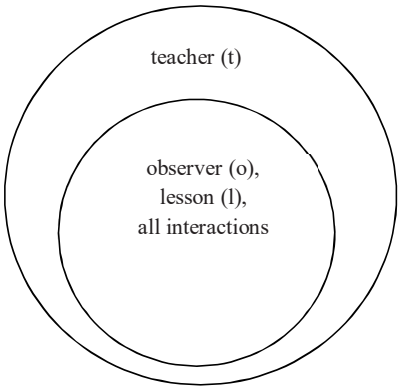
comparison to nested procedure, the crossed procedures are less flexible. The bias-confounded crossed procedure is more rigid, because it requires that teachers are grouped together and teachers within one group cannot simply be scheduled to visit lessons of any random colleague but only those of the few colleagues within his/her group. This restriction limits easy implementation in schools. However, it does provide additional information about possible sources of bias. As shown in the Venn diagram in Figure 5.3, if using a bias-confounded crossed procedure, it is possible to estimate bias due to teacher × observer interactions (facet ot) and to separate this bias from the bias due to observers (o). This is not possible in the bias-confounded nested procedure discussed next.

Bias-confounded nested procedure. The least complex procedure, or what we refer to as the bias-confounded nested procedure, has multiple observers visit one teacher's classrooms individually (see Figure 5.3). This procedure cannot indicate why classroom observations might emerge as unreliable. Rather, differences across lessons sum with differences among observers resulting in the "o, l, lo, to, tl, tlo, e". That is, all variance not attributable to differences in teaching are represented in a single error facet.

Figure 5.4

A schematic representation of the bias-confounded nested procedure (left) and the resulting variance decomposition (right).

	Observer					
	A	B	C	D	E	F
Teacher 1						
lesson 1	✓					
lesson 2		✓				
lesson 3			✓			
Teacher 2						
lesson 4				✓		
lesson 5					✓	
lesson 6						✓
Teacher *n*						



This procedure was examined by Kane et al. (2012) and advocated by Ho and Kane (2013, table 10). Its greatest benefit is its flexibility (anyone who receives training can perform a visit) in combination with an increased efficiency (requires fewer visits). Its greatest disadvantage is that the procedure provides no information about what specific actions can be taken in case where reliability is found too low. In our hypothetical example,

with three peers visiting three lessons, the procedure requires just 150 hours of observation instead of the 450 hours required by the previous two procedures. Also, schools do not have to find groups willing to visit multiple lesson taught by the same teacher together. Still, even this evaluation procedure demands considerable commitment from the school.

In summary, the crossed evaluation procedure, in which observers visit all the lessons together as a group (optimal situation from a psychometric perspective), is unrealistic for schools. Successful implementation instead requires a reduction of organizational complexity. The resulting situation is less than optimal, but more realistic, and it suffices to estimate the reliability of classroom observations. In this study, we have implemented the bias-confounded crossed procedure for this study.

5.3 Study aims and research questions

We explore the potential reliability of an evaluation design, as it has been implemented by actual schools. In so doing, we seek to replicate previous findings by Kane et al. (2012), Hill et al. (2012), and Ho and Kane (2013) that suggest that incorporating multiple lesson visits by multiple observers substantially increases reliability. This study also expands those previous works, by estimating the gains in reliability relative to certain absolute cutoffs (i.e., modest reliability $\boldsymbol{E}\rho^2 = .70$ and high reliability $\boldsymbol{E}\rho^2 = .90$) and explicitly relating the criteria to the different purposes of an evaluation, namely, formative feedback and summative decision, respectively.

Our focal research questions are as follows:

1. How many classroom observations by peers are required to achieve modest reliability and support formative feedback?
2. How many classroom observations by peers are required to achieve high reliability and support summative decisions?

5.3 Method

To investigate the research questions, peer observers in eight different schools across the Netherlands received training to perform observations of their colleagues. This type of collegial visitation fits the purpose of formative feedback, as well as current policies in the Netherlands (OCW, 2013a). The participating teachers each received three lesson visits by three different peers, after which we computed an evaluation score that could range from 0 to 31, such that 0 indicates the teacher poorly performed all of the teaching practices listed

in the instrument, and 31 indicated the teacher competently performed all of these practices. On the basis of this score, the teachers received feedback in a 20-minute, face-to-face conversation with the researcher, focused on their current teaching skills and the most likely options for improving their teaching.

5.3.1 Sample

Three different peers each observed a lesson taught by each teacher. The peers ensured that their lesson visits were scheduled for the same class. Using this procedure, we obtained 198 lesson observations of 69 teachers by 62 peers working at eight different schools across the Netherlands. The number of lesson observations is smaller than three times the number of teachers due to situational circumstances, such as when one of the three peers or the specific teacher was temporarily unavailable to perform or have lesson visits. Thus, 14 teachers were observed on only two occasions.

Teachers. Teacher experience ranged from 1 to 40 years (M = 13 years, SD = 10 years), and 62.1% of them were men. The non-representative gender distribution prompted us to check if male teachers might be evaluated differently than their female counterparts. An analysis of variance (ANOVA) revealed a negligible difference between male and female teachers ($F(1, 196) = 1.756$, $p = .18$). In addition, the teachers engaged in all available educational types: preparatory secondary vocational education (20.7%), senior general secondary education (46.5%), and university preparatory education (26.3%). The observed subjects were math (22%), history (21%), Dutch (20%), English (20%), and geography (4%), as well as German, Latin, economy, social sciences, science, religion, and construction (all ≤ 2%). Classroom observations took place between March and June 2014 and between February and June 2015.

Peer observers. Observers' teaching experience ranged from 1 to 40 years (M = 18 years, SD = 11 years), and 71.7% of them were males. Again, we checked if the unequal division of male and female teachers affected the overall evaluation results; the one-way ANOVA suggested no difference between male and female observers ($F(1, 196) = .01$, $p = .97$) or any indications of observer-gender × teacher-gender interactions ($F(1, 194) = .69$, $p = .56$). So, it seems likely that similar evaluation scores will be obtained in case that the division between males and females is more equal. In most instances, the peer observers were full-time teachers, though not all of them taught full-time. In modern Dutch schools, team managers frequently are part-time teachers, such that the boundaries between peer-

teacher and peer-manager are permeable. We use the word "peer" to refer to school personnel, all of whom have (previous) teaching experience.

5.3.2 Instrument

The International Comparative Analysis of Learning and Teaching is a Rasch-scaled observation instrument (Van de Grift, Helms-Lorenz & Maulana, 2014; Van der Lans, Van de Grift, & Van Veen, 2016). The most recent update of the instrument includes 31 items, each representing an effective teaching practice, such as "uses teaching methods that activate students." The items span six domains: safe learning climate, classroom management, clear instruction, activating students, teaching learning strategies, and differentiation (for details, see Van de Grift, 2014). Observers rated the items as either 0 = "insufficient" or 1 = "sufficient."

5.3.3 Procedures and training

The research procedure sought to simulate what a real implementation in schools would involve. That is, schools have limited time and resources for observation training, so for this study, the training lasted four hours, and observers were considered "limitedly trained." All colleague-teachers could participate in the training irrespective of their previous experiences with classroom observation. Also, we did not apply any tests or certification systems to prevent peer observers with insufficient inter-rater reliability from entering the classrooms; any peer who participated in the training was accepted as an observer, irrespective of his or her performance. These decisions are made because most schools have limited or no access to statistics, such that a real implementation would not involve the computation of inter-rater reliabilities. Also, schools are social organizations with their own group dynamics (Peterson, 2000). It is unlikely that they will (and can afford to) exclude willing peers from observing lessons. Therefore, this research aims to achieve sufficient reliability, given that schools decide to have all willing teachers participate in collegial visitation.

Observation training. The observation training involved a half-hour introduction to the instrument, after which the observers scored two lesson videos, each 20 minutes in length. Four different videos were available for the training, two in each training session. The videos were not randomly assigned; rather, in spring 2014, we used videos 1 and 2, and in spring 2015, we used videos 3 and 4. In both years, the training started with an easy

video followed by one that was more difficult to score. After each video, we calculated the percentages of observer agreement and discussed any problematic or confusing items. The videos of similar difficulty levels achieved similar consensus percentages: video 1 (74%) versus video 3 (75%) and video 2 (65%) versus video 4 (66%). Depending on the group, we also provided time to allow the trainees to express any insecurities about observing their peers.

5.3.4 Data preparation

During their observations, the peer observers were instructed to score as many items as possible. If a teaching practice was not observed, they had to decide whether in that lesson situation, the teacher should have used the practice, in which case the item was scored insufficient, or if the lesson situation did not allow for its performance, in which case the observers would leave the item blank. Of all item responses, only 3% were reported missing, so we considered them missing at random. We used procedures outlined by Raju, et al. (2006) to estimate an internal consistency coefficient similar to Cronbach's alpha. The internal consistency was high, $\rho_{(xx')} = .90$. However, consistency at the higher end of the measurement scale was considerably lower. Specifically, for raw scores of 30 and 31, the coefficient was less than $\rho_{(xx')} = .70$, so the evaluations did not consistently discriminate between the most excellent teachers.

5.3.5 Analysis

To examine the effect of adding additional peer observers, we used a Generalizability in Item Response Model (GIRT) methodology, as described by Briggs and Wilson (2007) and Choi (2013). The study design involves lessons (l) nested in observers (o) and teachers (t), crossed with items (i) (abbreviated (l:(o × t)) × i). The Venn diagram in Figure 5.5 is identical to the bias-confounded crossed procedure in Figure 5.3, except that it adds the item (i) facet, to describe the difference in chance on a positive score on the item describing the least complex teaching practice and the most complex teaching practice. This item facet is not a form of bias, because it describes a rank ordering in items identical for all teachers. In contrast, the facets item × observer (io) and item × teacher (it) should be interpreted as biases; they describe the degree to which the rank ordering is not identical for all teachers. For convenience, we refer to the facet of observer × teacher (ot), though more accurately,

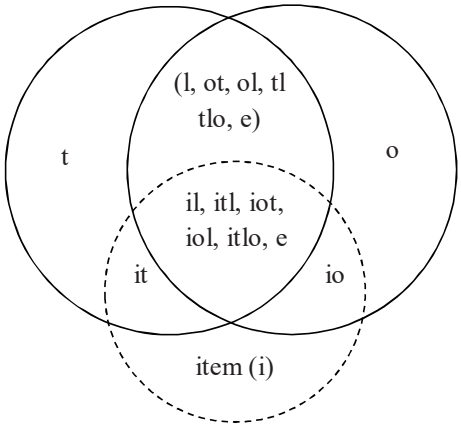
this facet is the sum of variation due to lessons (l), the observer × teacher interaction, and all interactions of facet lesson with the other facets.

To estimate the reliability coefficients, we used a two-step procedure. First, the generalizability (g-) study examines the amount of variation for which each facet, "t," "o," "i,", and their interactions can account. Second, the decision (d-) study examines the increase in reliability expected from adding more levels to a facet (Brennan, 2001). The Appendix F (Technical appendix) provides a more detailed explanation of the GIRT analysis.

G-study. The facets "o," "t," and "i" and their interactions were estimated using a multi-facet Rasch (1960) model, with the R package lme4 (Bates, et al., 2014). This package is a general statistical software package. Descriptions of how to formulate and estimate Rasch models using lme4 are available in de Boeck et al. (2011).

Figure 5.5

Venn diagram of the implemented bias-confounded crossed procedure with the item facet



D-study. The d-study examines the increase in reliability achieved by adding more peer observers. We studied two cut-off points, $\boldsymbol{E}\rho^2 = .70$ and $\boldsymbol{E}\rho^2 = .90$, and estimated how many observers would be required to achieve these levels. The logic underlying the d-study is that if the variance due to observers is large (e.g., 50% of total variance) and the number of observers is small, any particular observer adds considerably to the shifts across evaluation scores. Consequently, the relative weight of the observer facet (i.e., bias) should be greater, and the average evaluation score is unreliable. However, if the variance due to

observers remains similar, even with an increasing number of observers, the average evaluation score depends less on any particular observer. The relative weight of the observer facet then decreases, and the average evaluation score becomes more reliable. To estimate the relative increase in reliability with additional observers, a d-study assumes that the observer variance determined from the g-study is a true, unchanging reality, covering the complete range (or universe) of disagreement across classroom observers (Brennan, 2001). That is, this variance percentage can be expected with any number of observers. The d-study then varies the number of observers ($n_{(o)}$), thereby changing the relative weight of the observer facet in the reliability equation, to estimate the reliability levels with more or fewer observers. The d-study design is o × t × I. The capitalized "I" signifies that we consider the facet "items" as fixed, consistent with item response theory (Briggs & Wilson, 2007).

5.4 Results

To address the research questions, regarding how many classroom observations by limitedly trained peers are required to provide teachers with sufficiently reliable evaluations for the purposes of formative feedback ($\boldsymbol{E}\rho^2 \geq .70$) or summative decisions ($\boldsymbol{E}\rho^2 \geq .90$), we summarize the results of the G-study, with the design o × t × i, in Table 5.1.

Table 5.1

Variance Decomposition for the Multifacet Rasch Model

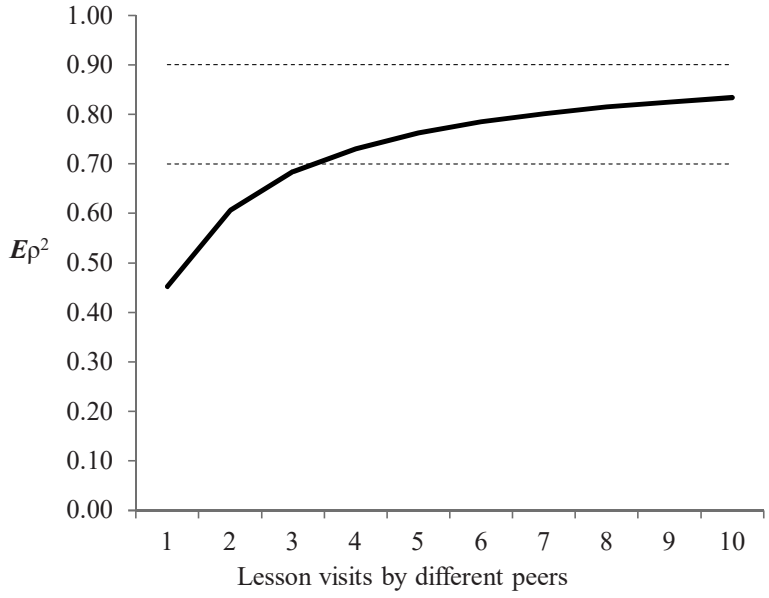
	$\boldsymbol{E}(\sigma^2)$	%
teacher (t)	1.29	0.22
observer (o)	0.37	0.06
item (i)	2.03	0.35
observer × teacher (l:ot)	1.07	0.19
item × teacher (it)	0.30	0.05
item × observer (io)	0.70	0.12
item × observer × teacher, e (i(ot:l), e)	.00	.00

Note. The number of estimated facets is different from the number published in Studies on Educational Evaluation. After the article was published we discovered that we had needlessly confounded two facets (observer and observer × teacher), which in fact could be separated.

As these results reveal, 22% of the variation in observed scores is due to true differences in teachers' skill. Furthermore, evaluations of the same teacher can vary substantially among observers: i.e. sum of facets o and ot. The variation due to observers is as great as the variation due to true differences in teaching skill. This substantial variation between observations—which, in the bias-confounded procedure, reflects the combined variance due to observers and lessons—is consistent with previous results (Hill et al., 2012; Ho & Kane, 2012; Kane et al., 2012). However, our results diverge in one important respect from previous findings: By using the GIRT method, we include the item (i) facet. This GIRT-based method includes more information for estimating evaluation scores than previous estimation techniques have, which should improve the reliability of the evaluation scores.

Figure 5.6

Expected increase in reliability ($\boldsymbol{E}\rho^2$) with increasing numbers of lesson visits by different peer observers.



Note. These estimates differ slightly from the results published in Studies on Educational Evaluation. They differ in terms of height, not in terms of direction. Resolving the confound (see note Table 5.1) also improved and changed the estimation of the teacher facet (t) and this decreased the estimated reliability coefficients by approximately .06.

This improvement is reflected in the expected reliability of an evaluation based on a single lesson visit, which is slightly higher than in previous works, yet still only $\boldsymbol{E}\rho^2 = .45$ (Figure 5.6). Figure 5.6 depicts how much this reliability is expected to increase with additional peer observers. To exceed the modest reliability criterion for formative feedback: i.e., reliability ≤ .70, a minimum of four lesson visits is required ($\boldsymbol{E}\rho^2 = .72$). The number of lesson visits required to exceed the high reliability criterion for summative decisions ($\boldsymbol{E}\rho^2 = .90$) exceeds 20 and it seems though not possible to reach this criterion if using only classroom observations. After 10 lesson observations ($\boldsymbol{E}\rho^2 = .83$) the relative increase in reliability of additional observations becomes negligibly small (i.e. <.01).

5.5 Conclusion and Discussion

This study investigates whether increasing the number of lesson visits and the number of peer observers also increases the reliability of teacher evaluation. Our findings indicate that reliable formative feedback demands observations of at least 4 different lessons by different peers, and reliable summative decisions demand that evaluators gather more than only classroom observations. These results align with previous findings that predict modest reliability when four different observers visit one another's lessons (e.g., Hill et al., 2012; Ho & Kane, 2013). This study further shows that this reliability also can be achieved with less complex evaluation procedures and without overly restrictive training protocols. After approximately 10 visits, additional classroom observations add almost negligible amounts of new information (increase in reliability less than .01). Hence, these values, of at least 4 and more than 10 visits, therefore are highly relevant for real-world evaluation practices by schools. They provide preliminary insights for how to start implementing classroom observations using cost-effective, manageable procedures, while still ensuring generally acceptable reliability.

The findings share similarities with results presented about five other classroom observation instruments in previous studies (Hill et al., 2012; Ho & Kane, 2013; Kane et al., 2012), including the classroom assessment scoring system (CLASS), the framework for teaching (FFT), the UTeach observation protocol (UTOP), the Mathematical Quality of Instruction (MQI), and the protocol for language arts teaching observation (PLATO) (see Table 5.2).

Table 5.2

Reliability indices reported for the ICALT in comparison with reliability indices reported for the FFT, CLASS, UTOP, MQI and PLATO (Kane, et al 2012, Table 11)

	one visit	Two visits	three visits	Four visits
ICALT	.45	.60	.68	.72
FFT	.37	.53	-	.67
CLASS	.31	.47	-	.63
UTOP	.30	.46	-	.63
MQI	.14	.24	-	.34
PLATO	.34	.50	-	.67

Therefore, the values of ≥ 4 (modest reliability) does not appear unique to the observation instrument that we applied; rather, it seems to be broadly characteristic of classroom observation instruments in general.

5.5.1 Alternative procedures to increase reliability

The number of lesson visits required to establish an acceptable reliability for summative evaluation is estimated as considerably more than 10 visits with each teacher. This currently seems an impossibly great number to achieve for schools and brings us to the question, What alternatives exist to increase the reliability of teacher evaluations? We discuss some possible directions, which should be subject to further research.

Kane et al. (2012) report that evaluations that combine different measures (e.g., student ratings, classroom observations, student achievement) are more reliable than evaluations based on classroom observations only. Such combinations accordingly might reduce the number of observers required. Alternatively, further development and improvement of the instrument we used could reduce these thresholds too. Our results suggest that classroom observations are currently biased by an item × teacher interaction and item × observer interaction (together 17% of the total variation). If this facet could be reduced to approximately 0%, the reliability will slightly increase but by no more than approximately .01. Finally, most previous studies in this field rely on procedures involving videotaped lessons (Hill et al., 2012; Ho & Kane, 2013; Kane et al., 2012). Videotaping technologies suggest some great potential for increasing flexibility, because the videos of teachers could be watched by observers at any time, so the observation hours could be

scheduled more flexibly. However, they also require schools to possess appropriate technical skills and equipment, particularly to ensure clear recordings of teachers' speech. The use of video also raises questions about whether these evaluations would be identical to evaluations based on actual lesson visits.

5.5.2 Limitations

This study has several limitations. First, the evaluation procedure (study design) did not incorporate differences across classes. A teacher's performance plausibly fluctuates from class to class, and the justification of summative decisions demands evidence of systematically poor or excellent performance across multiple classes, so the by Figure 5.6 estimated numbers of required lesson visits to achieve certain levels of reliability still are probably too low. Second, the current analysis estimates the increase in observation reliability for teachers with "average" teaching skill, to establish a single value of required visits. For performance at the extremes, generalizability theory instead predicts the need for fewer required observations (Brennan, 2001). Third, the terms modest and high reliability remain highly subjective. Although we use statistical cutoffs to define them, those very thresholds need to be subject to scrutiny and debate. Our criteria for reliability, following Nunnally (1978), have achieved wide acceptance. However, even Nunnally describes his criterion of .90 as a minimum to be tolerated and suggests that .95 should be the standard. Such a standard obviously would generate an even higher number of required lesson visits