University of Groningen

Once is not enough

van der Lans, Rikkert M.; van de Grift, Wim J.C.M.; van Veen, Klaas; Fokkens-Bruinsma, Marjon

Published in:
Studies in Educational Evaluation

DOI:
[10.1016/j.stueduc.2016.08.001](https://doi.org/10.1016/j.stueduc.2016.08.001)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2016

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

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 evaluative decisions based on classroom observation. *Studies in Educational Evaluation*, 50(September), 88-95.
<https://doi.org/10.1016/j.stueduc.2016.08.001>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations

Rikkert M. van der Lans*, Wim J.C.M. van de Grift, Klaas van Veen,
Marjon Fokkens-Bruinsma

Department of Teacher Education, Faculty of Social and Behavioral Sciences, University of Groningen, The Netherlands

ARTICLE INFO

Article history:

Received 4 February 2016

Received in revised form 15 July 2016

Accepted 3 August 2016

Available online 16 August 2016

Keywords:

Teacher evaluation

Classroom observation

Teacher feedback

Reliability

Generalizability Theory

Item Response Theory

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 ($E_p^2 \geq 0.70$) and that any summative decisions about their professional career have high reliability ($E_p^2 \geq 0.90$). A sample of 198 classroom observations by 62 colleagues of 69 teachers working at eight schools reveals that reliable feedback requires at least three lesson visits by three different observers and that reliable summative decisions require more than 10 visits. These findings mirror those reported through other observation instruments. This study accordingly offers directions for how schools can implement such procedures most cost-effectively.

© 2016 Elsevier Ltd. All rights reserved.

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; State of the States, 2013). In all of 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 given these aims signals that many countries are currently 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 because wrong decisions and feedback will harm individual teachers and will definitely not improve student learning outcomes.

Of these two purposes of teacher evaluation, the decisions regarding teacher selection currently receive the 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 to succeed in classrooms. This gives researchers and policymakers the moral obligation to carefully consider the reliability of their decisions. Clearly, evaluations

might be wrong if they were to select for dismissal those teachers who would have proven to be effective. Conversely, evaluations might be wrong if they were not selecting for dismissal those teachers who would have proven to be ineffective. Currently, priority is placed on attempting to avoid wrongly removing effective teachers, but this automatically leads to a situation in which ineffective teachers are wrongly retained (e.g., Winters & Cowen, 2014).

The provision of formative feedback has relatively less severe personal consequences. Nevertheless, feedback should also be based on a representative picture of the teacher's true teaching skill. In general, educational policies rely on classroom observations specifically for the purpose of targeting teachers who appear ineffective in some way and to provide them feedback (e.g., State of the States, 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 that 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 a few studies report on the reliability of these observation methods (e.g., Hill, Charalambous, & Kraft, 2012; Kane, Staiger, McCaffrey, Cantrell, Archer, & Buhayar, 2012). None of these studies relate reliability criteria to the two different purposes of teacher evaluation. This study seeks to determine whether classroom observations can achieve a reasonable level of reliability to support

* Corresponding author at: Department of Teacher Education, University of Groningen, PO Box 800, 9700 AV Groningen, The Netherlands.
E-mail address: r.m.van.der.lans@rug.nl (R.M. van der Lans).

both formative feedback and summative decisions, and if so, how many observations by how many separate observers are required to achieve this goal.

1. Theoretical background

1.1. Reliability and purpose of evaluation

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 are generally used for two different purposes. Therefore, different reliability criteria should apply to investigate whether instruments reliably support formative feedback and summative 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 were to 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 $E\rho^2 = 0.65$ 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 is currently determined by educational policies and school principals’ perceptions of what it takes to obtain 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 $E\rho^2 \geq 0.70$ suffices for formative feedback and for other instances in which the stakes are relatively low. Likewise, we suggest that a comparatively higher reliability level of $E\rho^2 \geq 0.90$ is the minimum criterion to use for summative decisions and for instances in which “a great deal hinges on the exact score made by a person on a test” (Nunnally, 1978, p. 245). Note that we use the notation $E\rho^2$ 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 E signifies that the reported coefficient reflects the expected reliability. It is the reliability we would expect if the evaluation procedure were to be repeated exactly.

1.2. 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). However, it is commonly acknowledged that one-time observations may be substantially biased by a bad moment or by a 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 implementing different classroom observation instruments, the findings are fairly consistent. Ho and Kane (2013) report reliability coefficients between 0.27 and 0.45, depending on the type of observer (teacher, peer or administrator). Kane et al. (2012) examine five classroom observation instruments and report coefficients of 0.37 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 0.37 and 0.46. That is, the reliability of single classroom observations is low and is generally less than 0.50. Previous works suggest that at least three lesson visits are required to achieve even modest reliability ($E\rho^2 \geq 0.70$) (Hill et al., 2012; Ho & Kane, 2013; Kane et al., 2012).

In addition to low reliability, the validity of one-time classroom visits has also been criticized on other grounds. One consideration is that if only one person is visiting it is clear that this is the person judging; therefore, observation scores cannot be anonymous (Scriven, 1981). This makes the appointed evaluator most vulnerable to criticism (French-Lazovik, 1981; Popham, 1988), which in turn provides an incentive to give lenient scores (Centra, 1975; Weisberg, Sexton, Mulhern, & Keeling, 2009). Weisberg, et al. stated that an evaluation procedure by which over 94% of the teachers observed are evaluated as performing sufficiently lacks validity. If multiple observers visit the classroom, then reporting the group average provides them some anonymity and protection.

1.3. 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. We 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.

1.3.1. Crossed procedure

This complex evaluation procedure requires a group of observers to visit all lessons together. An example of the crossed procedure is shown in Fig. 1. On the left side of Fig. 1, the evaluation procedure is visualized. Green check boxes reflect that the observer visited the lesson. On the right side of Fig. 1, the same evaluation procedure is visualized using a Venn diagram. Each circle in the Venn diagram is a facet. Each area 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 $\times$ teacher). In our notation, “e” refers to “error.” Furthermore, commas identify confounding facets. Confounding facets 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 Fig. 1 reflects that this part of the variation in scores may be explained by lesson $\times$ observer interactions, by teacher $\times$ lesson $\times$ observer interactions, and by measurement error. As such, this facet has no substantive interpretation.

Fig. 1. A schematic representation of the crossed evaluation procedure (left) and the resulting variance decomposition (right).

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 because the crossed design offers information about the reliability of the evaluation (i.e., true scores), 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. In this hypothetical case, if the group includes three observers, it would mean 450 h of lesson observation. Most schools lack the financial resources to hire external observers. Therefore, in practice, the observation group would likely consist 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 could be successfully implemented in schools even though this would be better from a psychometric point of view. In addition, visitation by an appointed group of appointed observers, will be vulnerable to criticism (French-Lazovik, 1981; Peterson & Chenoweth, 1992). In the crossed procedure, all teachers are evaluated by the same (small) group of observers, and because the observers might (over time) become more acquainted with certain subjects, or befriend some of their colleagues, it is likely that some of the teachers under evaluation will not feel that they are being treated equally. Note also that in a research setting, the strength of the crossed procedure is that it can take the resulting observer-teacher interactions into account but that in practice, there may be limited knowledge among some teachers about such statistical models. So, it is unlikely that schools can take adequate actions to

avoid tensions between teachers when implementing the crossed procedure.

1.3.2. Nested procedure

As a more flexible approach (Fig. 2), the nested procedure requires one group of observers to visit multiple lessons of one teacher together. The difference between the nested and the crossed procedure is that other teachers may be visited by other groups (see Fig. 2). This flexibility comes with a price, though. The procedure cannot reveal the extent to which reliability decreases due to observer $\times$ teacher interactions. Rather, the variance due to observer $\times$ 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 $\times$ teacher interactions.

None of the research referred to in this study has used the nested procedure. It offers benefits because it is more flexible with regards to who can perform the classroom observations in comparison to the crossed procedure. This flexibility is important because it provides opportunity for the careful selection of specific 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 no more efficient than the crossed procedure. Its implementation in our hypothetical, modestly sized school would require different groups of observers to visit 150 lessons together. Thus, if we again assume the groups include three peers, this procedure still demands 450 h of observation. In addition, despite that different groups may now perform the classroom observations, schools still have to schedule group visits. They need to find groups of observers who are willing to visit lessons together.

1.3.3. Bias-confounded nested procedure

The least complex procedure, what we refer to as the bias-confounded nested procedure, has multiple observers visiting

Fig. 2. A schematic representation of the nested evaluation procedure (left) and the resulting variance decomposition (right).

Fig. 3. A schematic representation of the bias-confounded procedure (left) and the resulting variance decomposition (right).

teachers' classrooms individually (see Fig. 3). This procedure cannot indicate why classroom observations might emerge as unreliable. Rather, differences across lessons sum with differences among observers, resulting in the facet "o, l, lo, to, tl, o, e". That is, all variance not attributable to differences in teaching is represented by a single facet of error.

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 (it requires fewer visits). Its greatest disadvantage is that the procedure provides no information about what specific actions can be taken in cases where reliability is found to be too low. In our hypothetical example, with three peers visiting three lessons, the procedure requires just 150 h of observation instead of the 450 h required by the previous two procedures. Additionally, schools do not have to find groups willing to together visit multiple lessons taught by the same teacher. Still, even this evaluation procedure demands considerable commitment from the school.

In summary, the crossed evaluation procedure, in which observers visit all of the lessons together as a group (an optimal situation, from a psychometric perspective), is unrealistic for schools. Successful implementation instead requires a reduction of organizational complexity, such that different observers visit lessons, and individual lesson visits are allowed. This situation is less than optimal, but it is more realistic, and it may suffice for estimation of reliability across classroom observations. Therefore, we implement a bias-confounded nested procedure for this study.

1.4. 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, both by estimating the gains in reliability relative to certain absolute cutoffs (i.e., modest reliability $E_p^2 = 0.70$ and high reliability $E_p^2 = 0.90$) and by explicitly relating the criteria to the different purposes of an evaluation, namely, formative feedback and summative decision, respectively. Accordingly, 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?

2. Method

To investigate the research questions, peer observers in eight different schools across the Netherlands received training in how

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, 2013). 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 behaviors listed in the instrument, and 31 indicated the teacher competently performed all of these behaviors. On the basis of this score, the teachers received feedback in a 20-min face-to-face conversation with the researcher. Feedback focused on educators' current teaching skills and the most likely options for improving their teaching.

2.1. 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 to have lesson visits. Thus, of the 69 teachers, 14 of them were observed on only two occasions.

2.1.1. 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 evaluations of male and female teachers ($F(1, 196) = 1.756$, $p = 0.18$). 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 $\leq 2\%$). Classroom observations took place between March and June 2014 and between February and June 2015.

2.1.2. 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 whether the unequal division of male and female teachers affected the overall evaluation results. The one-way ANOVA suggested neither any difference between male and female observers ($F(1, 196) = 0.01$, $p = 0.97$) nor any indications of observer-gender $\times$ teacher-gender interactions ($F(1, 194) = 0.69$, $p = 0.56$). Therefore, it seems likely that similar evaluation scores will be

obtained in the case of the division between males and females being 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.

2.2. 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 act, defined in terms 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, 2013). Observers rated the items as either 0 = “insufficient” or 1 = “sufficient.”

2.3. 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-based training. For this study, the training lasted four hours, after which point, observers were considered to be “limitedly trained.” All colleague-teachers could participate in the training irrespective of their previous experiences with classroom observation. Additionally, we did not apply any tests or certification screening 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 either limited or no access to staff skilled in statistical analytics, such that a real implementation would not involve the computation of inter-rater reliabilities. Additionally, schools are social organizations with their own group dynamics (Peterson, 2000). It is unlikely that they will (or could afford to) exclude willing peers from observing lessons. Therefore, this research aims to achieve sufficient reliability, given that schools typically decide to have all willing teachers participate in collegial visitation.

2.3.1. Observation training

The observation training involved a half-hour introduction to the instrument, after which the observers scored two lesson videos, each 20 min in length. Four different videos were available for the training, two per each training session. The videos were not randomly assigned. Rather, in the spring of 2014, we used videos 1 and 2, and in the spring 2015 semester, 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 insecurity about observing their peers.

2.4. Data preparation

During their observations, the peer observers were instructed to score as many items as possible. If a teaching behavior was not observed, they had to decide whether in that lesson situation, the teacher should have performed the behavior, in which case the

item was scored as 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; therefore, we considered them to be missing at random. We used procedures outlined by Raju, Price, Oshima, & Nering (2006) to estimate an internal consistency coefficient similar to Cronbach's alpha. The internal consistency was high, $\rho_{(xx')} = 0.90$. 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 $\rho_{(xx')} = 0.70$; therefore, the evaluations did not consistently discriminate between degrees of excellence among the most excellent teachers.

2.5. 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 Wislon (2007) and Choi (2012). The study design involves observers (o) nested in teachers (t), crossed with items (i) (abbreviated (o: t) $\times$ i). The Venn diagram in Fig. 4 is identical to the bias-confounded nested procedure in Fig. 3. The only difference is that it adds the item (i) facet to describe the difference in chance between scoring the item describing the least complex teaching behavior, and scoring the item describing the most complex teaching behavior. This item facet is not a form of bias because it describes a rank ordering of items identical for all teachers. By contrast, the facets item $\times$ observer and item $\times$ teacher 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 observers (o); though more accurately, this facet is the sum of variation due to observers (o), due to observer $\times$ teacher interactions (ot), and due to lessons (i), as we explained in the Background Theory section when describing the bias-confounded nested procedure we use.

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,” and “i,” can account. Second, the decision (d-) study examines the increase in reliability expected from adding more levels to a facet (Brennan, 2001).

2.5.1. G-study

The facets “o,” “t,” and “i” along with their interactions were estimated using a multi-facet Rasch (1960) model, with the R package lme4 (Bates, Maechler, Bolker, & Walker, 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, Bakker, Zwitser, Nivard, Abe, Tuerlinckx, and Partchev (2011).

Fig. 4. Venn diagram of the implemented bias-confounded nested procedure with the item facet.

2.5.2. D-study

The d-study examines the increase in reliability achieved by adding more peer observers. We studied two cut-off points, $Ep^2=0.70$ and $Ep^2=0.90$, 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 one 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 either more or fewer observers.

The d-study design is $(o: t) \times I$. The capitalized “I” signifies that we consider the facet “items” as fixed, consistent with item response theory (Briggs & Wilson, 2007).

3. 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 ($Ep^2 \geq 0.70$) or summative decisions ($Ep^2 \geq 0.90$), we summarize the results of the G-study, with the design $((o:t) \times i)$, in Table 1.

As these results reveal, 27% 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. The variation due to observers is almost 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, observer $\times$ teacher interactions, and lessons—is consistent with previous results (Hill et al., 2012; Ho & Kane, 2013; 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. This should improve the reliability of the evaluation scores.

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 is still only $Ep^2=0.51$ (Fig. 5). Fig. 5

Fig. 5. Expected increase in reliability (Ep^2) with increasing numbers of lesson visits by different peer observers (N_o).

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 ≤ 0.70 , a minimum of three lesson visits is required ($Ep^2=0.75$). To exceed the high reliability criterion for summative decisions, the minimum number of lessons required is more than 10 ($Ep^2=0.90$). Gathering data for more than 10 lesson observations would take more than a single school year. For our evaluation procedure, each teacher received three visits by different observers during one school year, which represented a realistic amount of visits per year. Thus, it would take about four years to gather the required 10 lesson observations.

4. Conclusions 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 three different lessons by different peers, and reliable summative decisions demand more than 10 different lessons observed by different peers. These results align with previous findings that predict modest reliability when three different observers visit each other's classrooms (e.g., Hill et al., 2012; Ho & Kane, 2013). This study further shows that this reliability can also be achieved with less complex evaluation procedures and without overly restrictive training protocols. These values of at least three visits for formative feedback and more than 10 visits for summative decision making, therefore, are highly relevant to schools' real-world evaluation practices. They provide preliminary insights for how to begin implementing classroom observations using cost-effective, manageable procedures, while still ensuring generally acceptable reliability.

The findings share some similarities with results presented by about five other classroom observation instruments in previous studies (Hill et al., 2012; Ho & Kane, 2013; Kane et al., 2012). These include 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). These previous studies report similar percentages of variance attributable to true differences in teaching skill (25–30%), as well as to observer bias and fluctuations across lessons (together, these two facets generally sum to 15–30%). Therefore, the values of ≥ 3 (modest reliability) and >10 (high reliability) observations do not

Table 1
Variance Decomposition for the Multifacet Rasch Model.

	$E(\sigma^2)$	%
teacher (t)	1.30	0.27
observer (o:t)	1.22	0.25
Item (i)	1.77	0.37
Item $\times$ teacher (i $\times$ t)	0.50	0.10
Item $\times$ observer, e ^a	0.00	0.00

^a Ime4 provides no estimate of the model residual (facet item $\times$ observer, e). In the d-study the facet is fixed at 0 and for completeness the Table reports this number. Fixing the residual to 0 is argued to be reasonable because the items fit Rasch model assumptions.

appear unique to the observation instrument that we applied. Rather, they seem broadly characteristic of classroom observation instruments in general.

4.1. Limitations

This study has several limitations. First, we determined that more than 10 lesson visits are required to attain the reliability criterion of $E_p^2 = 0.90$. However, 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; therefore, the estimated number of more than 10 observations is probably too low. Second, the current analysis estimates the increase in observational reliability for teachers with an "average" level of teaching skill, for establishing a single value of generally required visits. For performance at the extremes, however, generalizability theory instead predicts the need for fewer required observations (Brennan, 2001). Third, the terms 'modest reliability' and 'high reliability' remain highly subjective. Although we use statistical cutoffs to define them, those very thresholds need to remain subject to scrutiny and debate. Our criteria for reliability, following Nunnally (1978), have achieved wide acceptance. However, even Nunnally describes his criterion of 0.90 as a minimum to be tolerated and suggests that 0.95 should be the standard. Such a standard obviously would generate an even higher number of required lesson visits.

4.2. Alternative procedures to increase reliability

Undertaking more than 10 observations of each teacher might appear simply unrealistic in practice. Few schools have the organizational capacity to ensure that more than 10 different peers visit 10 different lessons by each teacher. This brings us to the following question: What alternatives exist to increase the reliability of teacher evaluations? We discuss some possible directions that could be the subjects of further research.

Kane et al. (2012) report that evaluations which combine different measures (e.g., student ratings, classroom observations, and student achievement) are more reliable than evaluations based on classroom observations alone. Accordingly, such combinations 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 $\times$ teacher interaction (10% of total variation). If this facet could be reduced to approximately 0%, the number of required lesson visits decreases slightly to more than 9. However, further research is required to explore the possible explanations of this interaction, to verify whether it is possible to reduce it. 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 videos also raises questions about whether these evaluations would be identical to evaluations based on actual lesson visits.

4.3. Implications for research

The numbers of classroom observations required to attain certain reliability criteria have important consequences for current

research. When studies use a single classroom observation, they can expect to report low to modest correlations at best because correlations become distorted by the overall low reliability of single classroom observations ($E_p^2 \leq 0.50$). In particular, the three lesson visits required to achieve modest reliability should be highly relevant to researchers interested in connecting classroom observations of teaching with other variables.

4.4. Implications for schools

These results suggest that those schools willing to invest in teacher evaluation should implement evaluation procedures that prescribe the use of different peers to visit each teacher's lessons, with a minimum of three visits. Only after three visits should schools provide feedback about what professionalization trajectories, courses, or schooling the teachers require to support their further development to avoid providing teachers with inaccurate feedback that potentially demotivates them. Furthermore, schools should gather at least 10 lesson visits before they use classroom observations to support summative decisions about tenure, salary, or dismissal. If schools collect fewer classroom observations, the chances that they wrongly offer tenure to or dismiss a particular teacher are unacceptably high.

4.5. Implications for peer-observers

In a strict interpretation of the evaluation procedure here proposed, this procedure denies individual peer-observers the opportunity give feedback because this feedback would lack reliability. This interpretation considerably constrains the learning opportunities for teachers. They are requested to invite peers into the classroom, after which the peers leave without giving teachers any feedback. Only after three peers have visited is feedback given. Clearly, this situation is undesirable because of a lack of direct feedback.

This would be, however, too strict of an interpretation of our results. Reliability concerns the degree to which scores can be generalized to other situations, and our findings indicate that evaluation outcomes based on one-time classroom observations do not sufficiently inform administrative stakeholders about the teacher's general teaching proficiency. Nevertheless, they do provide reliable insights about the specific lesson observed. Colleagues can provide direct feedback to teachers as long as they (1) do not rely on or mention specific item scores (to secure anonymity) and (2) give feedback about the specific lesson and avoid claims concerning the teachers' general teaching skill. Scriven (1981) mentions that classroom observation is unique in that it can give feedback about how the teacher reacted in a particular situation that may have occurred during the lesson, such as a misbehaving student or a particular student's questions.

4.6. Implications for teachers

We hope that explicitly stating these different reliability criteria provides teachers with some protection against a misapplication of evaluation instruments, especially if critical decisions about their careers are being made on the basis of single-lesson observations. Decisions based on this unreliable information may create an overall impression among teachers that a single poor performance in the classroom will have far-reaching consequences for their future employment. In turn, teachers likely would be reluctant to admit to any specific situations in which they feel incompetent, as Popham (1988), and Peterson (2000) caution. The criteria established in this study clearly demonstrate that one-time performance is not a sufficiently reliable means of supporting such decisions. With this information, teachers may have more

confidence both in demanding that their schools undertake additional observations, as well as in revealing performance gaps for which they seek to receive constructive feedback to improve their teaching, especially in those situations in which improvement is most desperately needed.

References

- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and Eigen. *R Package Version 1.1-7*. <http://CRAN.R-project.org/package=lme4>.
- Brennan, R. L. (2001). *Generalizability theory: statistics for social science and public policy*. New York: Springer-Verlag.
- Briggs, D. C., & Wilson, M. (2007). Generalizability in item response theory. *Journal of Educational Measurement*, 44, 131–155.
- Centra, J. A. (1975). Colleagues as raters of classroom instruction. *The Journal of Higher Education*, 46(3), 327–337. <http://dx.doi.org/10.2307/1980806>.
- Choi, J. (2012). Advances in combining generalizability theory and item response theory. *Doctoral dissertation*. Berkeley: University of California.
- Darling-Hammond, L., Amrein-Beardsley, A., Heafler, E., & Rothstein, J. (2012). Evaluation teacher evaluation. *Phi Delta Kappan*, 93, 8–15.
- De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Abe, H., Tuerlinckx, F., et al. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, 39, 1–25.
- DfEE (Department for Education), (2012). Teacher appraisal and capability. A model policy for schools. Retrieved June 30, 2015, from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/282598/Teacher_appraisal_and_capability.pdf.
- Firestone, W. A. (2014). Teacher evaluation policy and conflicting theories of motivation. *Educational Researcher*, 43, 100–107.
- French-Lazovik, G. (1981). Documentary evidence in the evaluation of teaching. In J. Millman (Ed.), *Handbook of teacher evaluation*. Beverly Hills, CA: Sage Publications.
- Hill, H., Charalambous, C. Y., & Kraft, M. A. (2012). When interrater-reliability is not enough: Teacher observation systems and a case for the generalizability theory. *Educational Researcher*, 41, 56–64. <http://dx.doi.org/10.3102/0013189X12437203>.
- Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Seattle, WA: Bill & Melinda Gates Foundation.
- Kane, T. J., Staiger, D. O., McCaffrey, D., Cantrell, S., Archer, J., Buhayar, S., et al. (2012). *Gathering feedback for teaching: combining high-quality observations with student surveys and achievement gains*. Seattle, WA: Bill & Melinda Gates Foundation.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17–64). 4th ed. Washington, DC: The National Council on Measurement in Education & the American Council on Education.
- Marzano, R. J. (2012). The two purposes of teacher evaluation. *Educational Leadership*, 70, 14–19.
- Mourshed, M., Chijioke, C., & Barber, M. (2010). *How the world's most improved school systems keep getting better*. London: McKinsey Company.
- Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation: An International Journal on Theory and Practice*, 12, 53–74.
- Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.
- OCW [Dutch Ministry of Education, Culture, Science] (2013). *Peer review in de praktijk [Peer review in practice]*. Rotterdam: VOION.
- Peterson, K. D., & Chenoweth, T. (1992). School teachers' control and involvement in their own evaluation. *Journal of Personnel Evaluation in Education*, 6(2), 177–189.
- Peterson, K. D. (2000). *Teacher evaluation: a comprehensive guide to new directions and practice*. Thousand Oaks, CA: Corwin press.
- Popham, W. J. (1988). The dysfunctional marriage of formative and summative teacher evaluation. *Journal of Personnel Evaluation in Education*, 1, 269–273.
- Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2006). Standardized conditional SEM: a case for conditional reliability. *Applied Psychological Measurement*, 30, 1–12. <http://dx.doi.org/10.1177/0146621606291569>.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielsen & Lydiche.
- Scriven, M. (1981). Summative teacher evaluation. In J. Millman (Ed.), *Handbook of teacher evaluation*. Beverly Hills, CA: Sage Publications.
- Shavelson, R., & Dempsey-Atwood, N. (1976). Generalizability of teaching behavior. *Review of Educational Research*, 46, 553–611.
- State of the States (2013). *Connect the dots: Using evaluations of teaching effectiveness to inform policy and practice*. Washington, DC: National Council on Teacher Quality.
- Strong, M. (2011). *The highly qualified teacher: what is teacher quality and how do we measure it?* NY, New York: Teachers College Press.
- Toch, T., & Rothman, R. (2008). *Rush to judgment: teacher evaluation in public schools*. Washington DC: Education Sector.
- van de Grift, W. J. C. M. (2013). Measuring teaching quality in several European countries. *School effectiveness and school improvement*, 25(3). <http://dx.doi.org/10.1080/09243453.2013.794845>.
- van de Grift, W. J. C. M., Helms-Lorenz, M., & Maulana, R. (2014). Teaching skills of student teachers: Calibration of an evaluation instrument and its value in predicting student academic engagement. *Studies in Educational Evaluation*, 43, 150–159. <http://dx.doi.org/10.1016/j.stueduc.2014.09.003>.
- van der Lans, van de Grift, & van Veen (2016). Developing an instrument for teacher feedback: Using the Rasch model to explore teachers' development of effective teaching strategies and behaviors. Manuscript submitted for publication.
- Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., et al. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. *New Teacher Project*.
- Winters, M. A., & Cowen, J. M. (2014). Who would stay, who would be dismissed? An empirical consideration of value-added teacher retention policies. *Educational Researcher*, 42, 330–337. <http://dx.doi.org/10.3102/0013189X13496145>.