Studies in Educational Evaluation 50 (2016) 88–95

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

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 ($E\rho^2 \geq 0.70$) and that any summative decisions about their professional career have high reliability ($E\rho^2 \geq 0.90$). 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.

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).

http://dx.doi.org/10.1016/j.stueduc.2016.08.001
0191-491X/

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 $\boldsymbol{E}\rho^2 = 0.65$ 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 $\boldsymbol{E}\rho^2 \geq 0.70$ 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 $\boldsymbol{E}\rho^2 \geq 0.90$ 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 $\boldsymbol{E}\rho^2$ 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 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 ($\boldsymbol{E}\rho^2 \geq 0.70$) (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 × 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 × observer interactions, by teacher × lesson × observer interactions, and by measurement error. As such, this facet has no substantive interpretation.