university of groningen / faculty of behavioural and social sciences / teacher education

Once is not enough:
Establishing reliability criteria for teacher evaluation based on classroom observations

Rikkert van der Lans, Wim van de Grift, & Klaas van Veen

Correspondence: r.m.van.der.lans@rug.nl

Abstract

Classroom observation is the most implemented method to evaluate teaching. To ensure reliability, researchers often train observers extensively. However, schools have limited resources to train observers and often lesson observation is performed by limitedly trained or untrained colleagues.

In this study an evaluation procedure is implemented which is dependent on classroom observation by limitedly trained (3hours) peers. To study whether observations have sufficient reliability, two different criteria are specified: one more lenient for formative evaluation and the other one strict for summative evaluation. The study aims to explore whether these criteria are realistic for schools.

The sample contains 198 lesson observations of 69 teachers, by 62 peer-colleagues. Two different aspects of reliability are studied: (1) generalizability: the extend to which other lessons observed by other peers would give the same evaluation result, (2) person fit: the extend to which observations fit model assumptions. The results show that three peer-observers are required to achieve sufficient reliability for formative purposes, while more than 10 are required to achieve sufficient reliability.

Background and definitions

Generalizability is the degree to which another new observation of a lesson taught by the specific teacher to that specific class would give the same result.

Person fit is the degree to which the specific observation for the specific teacher fits model presumptions. The most fundamental model assumptions are:

1. Effective teaching methods and strategies can be grouped in six domains
2. These six domains can be ordered cumulatively (See Figure 1).

Previous work has confirmed the cumulative ordering is plausible (e.g., van de Grift, et al. 2014)

Misfit is identified by unexpected item scorings (see Figure 1). Teacher 3 has an error, since the model expects items in the domain climate to be scored 'correct' if items in subsequent domains are scored correct.

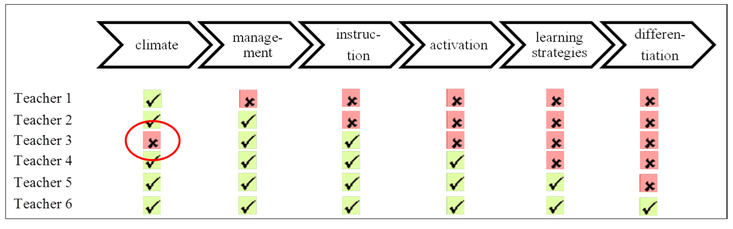

Figure 1. The six domains cumulatively ordered. All but one item response fits.

Criteria for formative feedback

1. Reliability of **≥.70** suffices for formative feedback
2. **≥ 70%** of the teacher's scores follow the cumulative ordering

Criteria for summative decisions

1. Reliability of **≥ .90** is the minimum criterion
2. **≥ 90%** of the teacher's scores follow the cumulative ordering

Research Question 1:

How reliable are lesson observations of limitedly trained colleague teachers for formative feedback and summative judgments?

Method

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 peer-observations of 69 teachers, by 62 peer observers.

Instrument. The International Comparative Analysis of Learning and Teaching (ICALT) is a Rasch-scaled observation instrument counting 31 items (van de Grift, et al. 2014). Items represent an effective teaching method or strategy and span six domains: safe learning climate, classroom management, clear instruction, activating students, teaching learning strategies, and differentiation. Observers rated the items as either 0 = "insufficient" or 1 = "sufficient."

Design. Peer-observations (o) are nested in teachers (t). Also, since every peer observed another lesson, the differences between lessons (l) are confounded with differences between peers. The resulting design is abbreviated with (o:t, l). This design has been recommended by Ho and Kane (2013, table 10), because it minimizes the demands for schools, while it should provide acceptable estimations of reliabilty.

Results research question 1:

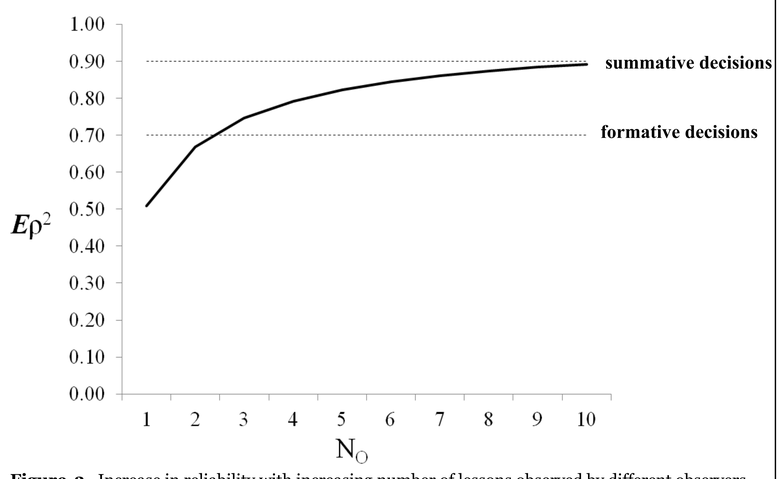

Figure 2. Increase in reliability with increasing number of lessons observed by different observers. $E\rho^2$ = reliability, N_O = number of classroom observations.

Conclusions

1. Providing teachers with reliable feedback requires three lesson visits by three different observers
2. Ensuring reliability of summative decisions requires more than ten lessons visits by different observers

Discussion topic: **is this evaluation procedure realistic for schools?**

Research question 2:

What percentage of lesson observations shows sufficient person fit for formative feedback, and what for summative judgments?

Method

Analysis . Using eRm (Mair & Hatzinger, 2007) Rasch analysis is used to estimate the cumulative ordering in effective teaching methods and strategies (see handout). It is verified whether the ordering is consistent with those reported in other research.

A Guttman simplex factor analysis is performed to test fit of one-dimensionality of the cumulative pattern. Using Circum software (developed by Browne, 1991) results suggest modest fit (RMSEA = .08).

Then, $G_{N\,ORMED}$ (Meijer, 1996) is used to give an indication of the percentage of deviations from the cumulative ordering. $G_{N\,ORMED}$ is based on the number of Guttman errors (see Figure 1) and it divides the number of Guttman errors by the total possible number of Guttman errors. The resulting statistic reflects the percentage of deviations from the cumulative ordering.

Results research question 2:

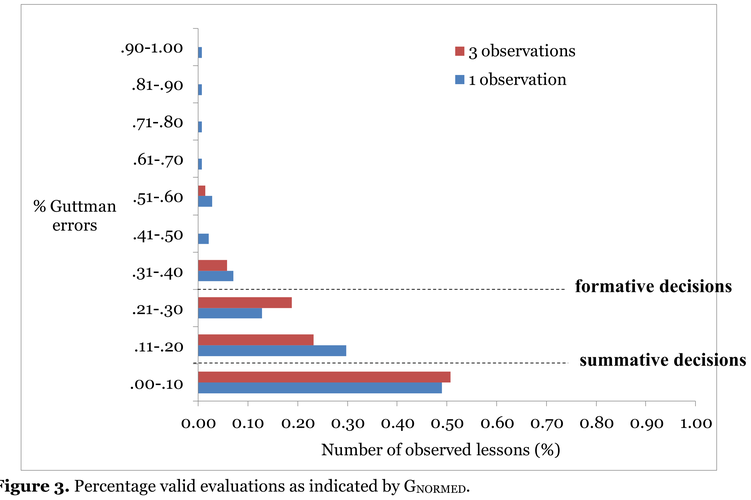

Figure 3. Percentage valid evaluations as indicated by G_{NORMED}.

Conclusions

1. Single lesson observations more frequently deviate from model expectations. For 15% of the teachers model estimations are incorrect
2. Ensuring more lesson visits reduces the number of deviating evaluations. In case of three lesson observations (by different peers) only for 8% of the teachers model estimations are incorrect

Discussion topic: **How is G_{NORMED} useful for teacher evaluation?**