Abstract

Classroom observation is the most implemented method to evaluate teaching. To ensure reliability, researchers often train teachers 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 (jihous) 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).

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.

Person fit
misfit

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

Results research question 1:

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?

Results research question 2:

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 GORMED useful for teacher evaluation?