Studies in Educational Evaluation 58 (2018) 112–121

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/locate/stueduc

The use of secondary school student ratings of their teacher's skillfulness for low-stake assessment and high-stake evaluation

Rikkert M. van der Lans*, Ridwan Maulana

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

ARTICLE INFO

Keywords:
Teacher evaluation
Teacher assessment
Student ratings
SET
Reliability
Generalizability theory

ABSTRACT

Previous studies in higher education have shown that the reliability of student ratings of teaching skill increases if multiple ratings by different students are aggregated. This study examines the generalizability of these findings to the context of secondary education. Also, it seeks to validate these findings by comparing reliability levels estimated by the routinely used nested design with those estimated using a more complex design. The sample consisted of 410 students from 17 classes rating 63 teachers working at eight schools across the Netherlands. Using the nested design, the study replicates findings of previous studies in higher education. The findings illustrate how the reliability level of secondary school students' ratings increases with an increasing number of students. However, these replicated reliability levels were not validated by the more complex design which provided lower estimates. This indicates that the nested design may not provide accurate estimations of rating reliability.

1. Introduction

This study examines the reliability of student ratings of teachers' classroom teaching in secondary education using generalizability theory (Cronbach, Gleser, Rajaratnam, & Nanda, 1972). Generalizability theory has been applied in the context of higher education by Kane, Gillmore, and Crooks (1976) and Gillmore, Kane, and Naccarato (1978). In addition, some other studies in higher education report between year and/or between-class correlations (e.g., Feistauer & Richter, 2016; Marsh, 1982; Marsh & Hocevar, 1991). Though formally these studies are not "true" generalizability studies, they align with its general principles. Together these works continue to dominate the discourse about reliability of student ratings which can be illustrated by their mentioning in reviews by Benton and Cashin (2012); Marsh (2007), and Richardson (2005).

The literature on student rating reliability still is much thinner for secondary and primary education, though some studies have addressed the topic (e.g., Bill & Melinda Gates Foundation, 2012; Fauth, Decristan, Rieser, Klieme, & Büttner, 2014; Lüdtke, Trautwein, Kunter, & Baumert, 2006; Peterson, Wahlquist, & Bone, 2000; Polikoff, 2015; Panayiotou et al., 2014). However, none of the previous studies applied generalizability theory. By performing a generalizability study in the context of secondary education this study aims to foster further understanding of the reliability of secondary school student ratings.

An additional advantage of the application of the generalizability theory is that it provides the possibility to explore whether current knowledge about reliability of (secondary school) student ratings depends on the design of the study. The role of the research design remains an underrepresented topic in studies on the reliability of student ratings. Previous research has routinely applied the nested research design in which one class of students rates their teacher and another class of students rates another teacher (e.g., Fauth et al., 2014; Kane et al., 1976; Gillmore et al., 1978; Lüdtke et al., 2006; Polikoff, 2015[1]) and this has made some to doubt the accuracy of previous estimations of reliability (e.g., Morley, 2012).

Our study has two aims: first it attempts to replicate previous findings in higher education of the reliability presented by Kane et al. (1976) and Gillmore et al. (1978) and summarized by Marsh (2007) in the context of secondary education. In specific it is examined whether Marsh's claim that approximately one class consisting of 25 students is required to achieve a reliability level of ≥ 0.90 generalizes to the context of secondary education. The second aim of the study is to examine whether the estimated reliability based on the nested design in which one class rates one teacher (in the subsequent text also referred to as one-class-one-teacher design) is validated by the more complex half block design in which one class rates multiple teachers (in the

* 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).

[1] We position Polikoff (2015) in this list despite that his sample includes ratings from multiple school-years and in which teachers might have switched classes. The reason is Polikoff's data analysis strategy, which regresses the total score of year 1 on the score of year 2. This strategy considers the ratings of each year to be fixed.

https://doi.org/10.1016/j.stueduc.2018.06.003
Received 27 March 2017; Received in revised form 5 June 2018; Accepted 8 June 2018

subsequent text also referred to as the: one-class-multiple-teacher design). This part of the study seeks to validate findings based on the nested design.

2. Background

This study examines reliability of teachers rated by secondary school students because they are (potentially) used for teacher evaluation and teacher assessment purposes. In the study, the term "evaluation" refers to the specific application of student ratings of their teachers' teaching skill to inform "high-stake" decisions. We are aware of the additional connotation of the term evaluation in the general literature with formative purposes (e.g., Bill & Melinda Gates Foundation, 2012; Marzano & Toth, 2013). However, in our view using the term "evaluation" for both the summative purpose of "high-stake" decisions and formative purposes of feedback and coaching should be avoided to prevent confusion in the field. The reason for this is that the requirements to be met for summative and formative "evaluations" differ. Therefore, we propose to disentangle the general use of the term evaluation by restricting it to refer to summative purposes and to use the term "assessment" for formative purposes.

2.1. Reliability: a criterion for valid use

This study approaches reliability as evidence supporting the validity for using scores for specific purposes (Kane, 2013). This approach is consistent with other studies: for example, Ho and Kane (2013) suggest that a reliability of 0.65 is required to use classroom observation scores for certain evaluation and assessment purposes and Nunnally (1978) suggested that reliability of 0.70 is minimally required to use data for low-stake explorative research purposes, whereas a reliability of 0.90 is minimally required if decisions have personal consequences.

We connect these criteria to the two purposes of evaluation and assessment. Teacher evaluation involves summative decisions concerning tenure, salary, and dismissal which can affect personal lives, while the teacher assessment concerns advice for improvement and training intended to affect professional practice only. Because of this, we propose that a reliability level of 0.90 is required if intentions are to use the obtained information in support of high-stake teacher evaluation, whereas a reliability level of 0.70 might be considered sufficient if intentions are to use the obtained information in support of (lower-stake) teacher assessment.

Additionally, the literature concerning (teacher) evaluation and assessment distinguishes between two approaches, namely norm-referenced and criterion-referenced approaches (e.g., Brennan, 2001; Lok, McNaught, & Young, 2016). In a norm-referenced approach, teachers' scores are compared to other teachers' scores and a predetermined percentage of teachers would obtain a certain qualification (e.g., "low", "average", or "high"). A potential disadvantage of this approach is that it may lead to improper decisions because if all teachers are highly skilled then still a predetermined number of teachers would obtain the qualification "low" regardless of their absolute performance (Lok et al., 2016). In the criterion-referenced approach, teachers' scores are compared to some absolute standard to obtain a certain qualification (e.g., "below", "similar", or "above the standard"). A potential disadvantage of this approach is that it may prompt assessors and evaluators to bias their scores upwards to ensure that teachers reach the criterion (Lok et al., 2016; Weisberg et al., 2009).

Generalizability theory provides two operationalizations of reliability: (1) the generalizability coefficient (ρ) and (2) the index of dependability (ϕ) (Brennan, 2001; Kane & Brennan, 1977; Wiley, Webb, & Shavelson, 2013). The generalizability coefficient examines the relative consistency in the rank ordering of teachers' scores. It can provide evidence supporting the validity to give a norm-referenced interpretation to evaluation or assessment outcomes (Brennan, 2001; Wiley et al., 2013). The index of dependability examines the absolute deviations from teachers' scores. It may provide evidence supporting the validity to give a criterion-referenced interpretation to evaluation or assessment outcomes (Brennan, 2001; Wiley et al., 2013). The current study operationalizes reliability as the index of dependability and, thus, results may support the validity for using scores in a criterion-referenced approach.

2.2. Prior evidence of reliability of student ratings

In this study, reliability is conceptualized in line with generalizability theory as the dependability of scores on the teachers' teaching skill (Brennan, 2001). Dependability is the extent to which scores inform about teaching skill. Generalizability theory provides an understanding about how (dis)aggregation of scores will change their dependability. For example, Marsh (2007) reviews that the correlation of ratings by two randomly chosen students usually is in the 0.20′s, whereas if these student ratings are aggregated into class average ratings by 25 students or more their correlation may exceed 0.90. Thus, the dependability of a single student rating on the teachers teaching skill is low, whereas the dependability of the class means is large (Kane & Brennan, 1977). Generalizability theory has been applied in previous studies in higher education (Gillmore et al., 1978; Kane et al., 1976). Because the application of generalizability theory remains underrepresented in secondary education, we will use these studies from higher education to get some indications about what might be expected in the present study.

Kane et al. and Gillmore et al. compared different combinations of teachers and courses to verify whether student ratings are more dependent on the teacher than on the course taught. They report that reliability is mainly affected by the number of students, and much less by the item content and on the subject course taught. Subsequent correlational studies by Marsh (1982) and Rindermann and Schofield (2001) broadly corroborated these findings. Feistauer and Richter (2016) report that the size of variance components (or facets) – from which reliability coefficients are generally estimated – may vary between subscales and courses, but also their results indicate that student ratings are mainly dependent on the number of students. In summary, research suggests that there are various factors affecting the rating reliability, but there is a general consensus that the number of students is a dominant factor affecting the rating reliability.

The application of generalizability theory allows for comparison with the above lines of research. However, the choice to use generalizability theory also complicates comparison with other studies examining reliability of student ratings, including Polikoff (2015); Fauth et al. (2014); Panayiotou et al. (2014) and to some extent Lüdtke et al. (2006). Polikoff (2015) recently addressed the year-to-year stability of student ratings and reports fixed regression weights. It is not straightforward how to compare these regression weights with the reliability coefficients studied here. Fauth et al. (2014) and Panayiotou et al. (2014) study the validity of student ratings using structural equation models. The model fit indices they report may be perceived as providing information about the reliability of student ratings, but these also are complex to compare with the here applied generalizability coefficients. Finally, Lüdtke et al. (2006) compare various statistical approaches to estimate reliability most of which are difficult to compare to the here studied generalizability coefficients. Exceptions are the intra-class correlations (ICC) and ICC(2). The latter ICC(2) extends the regular ICC equation with the Spearman-Browne prophecy (Lüdtke et al., 2006). The ICC(2) overlaps with the generalizability coefficient of the nested design that is studied in the present study (Brennan, 2001).

2.3. Validating the evidence of reliability of student ratings

Nearly all studies on the reliability of student ratings (e.g., Gillmore et al., 1978; Kane et al., 1976; Lüdtke et al., 2006; Marsh, 2007) make use of the same nested one-class-one-teacher design. The nested design