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

van der Lans, Rikkert M.; Maulana, Ridwan

Published in:
Studies in Educational Evaluation

DOI:
10.1016/j.stueduc.2018.06.003

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 is appealing because in terms of time and resources it is the most efficient design. However, according to generalizability theory its application is valid only if it can be shown a more complex design results in comparable reliability estimates (i.e. $\phi_{(\text{nested design})} = \phi_{(\text{more complex design})}$) (Brennan, 2001; Cronbach et al., 1972; Shavelson & Webb, 1991). For the specific case of student ratings, a study comparing the nested design with the most complex completely crossed design would examine the validity of two assumptions: (1) class mean ratings are dependent on teachers' teaching skill and independent of class composition and (2) class mean ratings are independent of the specific teacher-class combination. If the first assumption is violated, then some classes rate all their teachers higher or lower compared to other classes. If the second assumption is violated, then the specific class rates their teacher higher or lower than other classes while in general the specific class does not rate teachers more or less favorably compared to other classes. This interaction effect would hint that the class may have a bias towards that specific teacher (see Brennan (2001) and Shavelson and Webb (1991) for further descriptions of assumptions of nested and confounded designs).

Some limited empirical evidence concerning the validity of these assumptions is found in Marsh and Hocevar (1991) who report the longitudinal stability of ratings by 13 different classes of higher education students. Because teachers taught each year another class, the study by Marsh and Hovecar might be perceived as operationalizing the situation where teachers are rated by different classes. They conclude that, on average, teachers receive fairly stable student ratings by these different classes across 13 years, but note that the group average hides the individual differences observed among teachers. They explore this individual variation using separate regression analyses for each individual teacher and report approximately 16% between-class variation in student ratings of the same teacher over the school-years. Because the design of Marsh and Hocevar (1991) confounds variation between school-years with variation between classes it would be invalid to conclude that the between class differences are approximately 16%. Teachers may have developed across the years. The result allows for the conclusion that the between-class variation will not have been larger than 16%, but because some of this variation will be due to differences in teaching skill between school years, the percentage of between-class variation was probably lower.

This study implements a half block design in which one class rates multiple teachers. The half block design is more complex than the nested design. The design can address the first of the two assumptions. To our knowledge, no current studies have examined a design in which multiple teachers are rated by the same class. Therefore, the second aim of this research is to examine whether the reliability coefficients provided by the nested design can be validated when implementing a more complex design.

2.4. Research questions and hypotheses

Given the above, the study will address the following two research questions:

1 To what extend is the reliability of ratings by students in secondary education similar to the reliability of ratings by higher education students as reported by Kane et al. (1976) and Gillmore et al. (1978) and when using the nested one-class-one-teacher design?

Despite differences in structure and characteristics between higher education and secondary education, Kane et al. (1976) mentioned that students' ratings reliability is mainly dependent on the number of students, suggesting that other context variables might not matter so much for reliability. Hence, we hypothesize that the reliability of ratings previously found in higher education will be replicated in the secondary education context.

2 Are the findings provided by the nested one-class-one-teacher design validated by the more complex one-class-multiple-teacher design?

In the light of previous results presented in Marsh and Hocevar (1991) it is expected that the more complex design leads to different estimates of reliability compared to findings obtained with the nested design. However, because we have inadequate evidence supporting that the one-class-multiple-teacher design will lead to lower of higher estimations of ratings reliability, we will not formulate a hypothesis and leave this as an exploratory question.

3. Method

3.1. Sample

The sample consisted of 410 students from 17 classes rating 63 teachers working at eight schools across the Netherlands. Of the students 46.5% are boys, age varied between 12 and 16, but over 90% of the students were between 13 and 15 years old. Class size varied from 16 to 31 students, with an average number of 25 students per class.

Per class four teachers participated, which were all rated by the students in that class. Not all schools succeeded to complete the project. Of the 17 classes, five classes of students rated only three teachers. Participating teacher quartets could teach varying subjects, but schools were strongly advised to select teachers teaching Dutch, English (as a foreign language) history, and math and 80.8% of the teachers taught these subjects. Other subjects included in the sample were: economy, geography, social sciences, religion, physics, and technical drawing and construction (see Table 1).

Teacher experience ranged from 0 to 40 years of experience. Of the participating teachers 60.5% are male. The unequal distribution in teacher gender prompted us to check whether gender could affect the study results. Outcomes could range from 0 to 40 points and the outcome was included as dependent variable in a multilevel mixed model including a random intercept for teacher and class and a fixed effect for teacher gender and student gender. The analysis indicated no significant difference between male ($M = 28.35$) and female teachers ($M = 29.51$) ($F(1, 55.82) = .26$, $p = .61$), but it did indicate a significant interaction between student gender and teacher gender ($F(1, 1364.22) = 4.78$, $p = .029$). While girls evaluate male (29.72) and female (29.31) teachers similar, boys are found to rate male teachers (28.82) slightly higher than female teachers (27.89). However, given the small differences in average ratings, it seems reasonable to argue that the unequal distribution in gender has only small implications for the study results.

3.2. Instrument

The "My Teacher" questionnaire consists of 40 items describing various teaching practices which are proven to increase student learning (Maulana, Helms-Lorenz & van de Grift, 2015; van der Lans,

Table 1
Overview of the subjects included in the sample.

Subject	$n_{(\text{teachers})}$	%
English	14	22.2
History	12	19.0
Dutch	13	20.6
Mathematics	12	19.0
Geography	4	6.3
Religion	2	3.2
Social sciences	2	3.2
Physics	2	3.2
Economy	1	1.6
Technical drawing and construction	1	1.6

Table 2
The six domains and corresponding items of the "My Teacher" questionnaire.

Domain	Example item	Answer option	
	My teacher…	*rarely*	*often*
Safe learning climate	… ensures that I feel relaxed in class	0	1
Efficient classroom management	… applies clear rules	0	1
clear and structured explanation	… clear instruction uses clear examples	0	1
Activating teaching methods	… involves me in the lesson	0	1
Teaching learning strategies	… explains how I should study something	0	1
Differentiation	… knows what I find difficult.	0	1

van de Grift & van Veen, 2015). Items in the questionnaire pertain to six domains (see also Table 2), specifically: safe learning climate (SLC), efficient classroom management (ECM), clear and structured explanation (CSE), activating teaching methods (ATM), teaching learning strategies (TLS), and differentiation in instruction (DII) (sometimes referred to as adaptation of instruction). Students rated items on a two-point scale, specifying whether their teachers performed the teaching practice "rarely" or "often".

Previous research applied Rasch (1960) analysis to investigate the internal structure of the questionnaire (Maulana et al., 2015; van der Lans et al., 2015). These studies suggest that most of the 40 items fit the Rasch model assumptions and relate to one single cumulative dimension. (Fig. 1). The cumulative ordering provides evidence that teaching practices may be interpreted in terms of complexity. In this interpretation, teaching practices related to domains at the right-side of Fig. 1 are more complex, because they are only scored in combination with teaching practices at the left-side of Fig. 1. Thus, according to the model in Fig. 1, more effective teaching requires teachers to perform more teaching practices simultaneously.

Broadly, previous studies support that student ratings on the "My Teacher" questionnaire follow a scoring pattern as illustrated in Fig. 1 (Maulana et al., 2015; van der Lans et al., 2015).

3.3. Questionnaire protocol

The design connected one class to four teachers. Students within this class completed the same questionnaire each time concerning one of the four participating teachers. If within one school multiple classes participated each class rated a unique group of four different teachers to prevent cross-classification of teachers.

Schools administered the questionnaires themselves so that the data gathering increases the ecological validity of our results because it closely reflects how schools may implement the design. Schools were given freedom in assigning the classes to teachers and choosing the moment and place of questionnaire administration. Schools were not allowed to administer more than two questionnaires in one lesson hour to provide students sufficient time to complete the questionnaires and to prevent fatigue. Schools and teachers were not allowed to have students complete the questionnaires at home.

3.4. Data preparation

3.4.1. Missing responses

Of the total number of 58,400 item responses only 1.3% reports a missing value. This number of missing values is acceptable and not expected to substantially distort the results. However, some specific questionnaires report a considerable number of missing values. It was decided to exclude 32 (2.1%) questionnaires which counted less than 35 valid item responses. These 32 questionnaires correspond to 22 students. Due to this, the available number of questionnaires decreased slightly from 1445 to 1413.

3.4.2. Missing cases

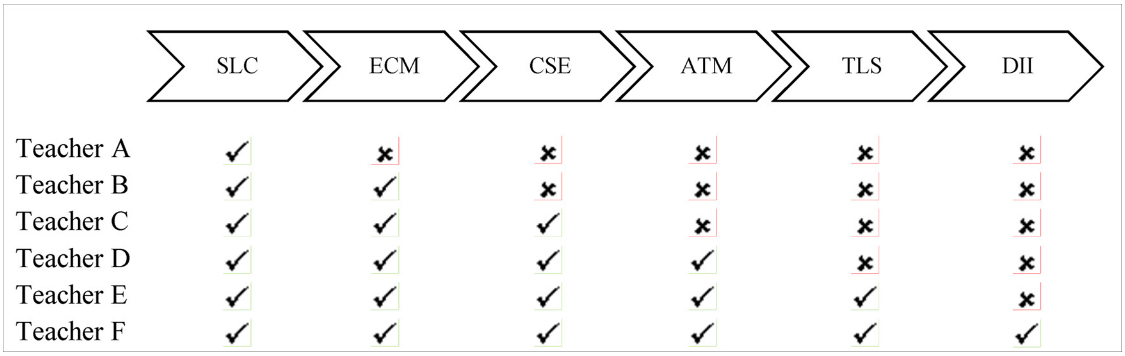
Of the 410 students 29 (7.1%) are coded as "missing cases" because they failed to return more than two questionnaires or because they returned more than two questionnaires having too many missing values. These students were omitted from the reliability analysis, though they provided at least one valid questionnaire. The reason for this decision is that this group adds to the teacher variance, but – since they only contributed one or two valid questionnaires – add no or only few class and student variance. Missing cases are equally distributed across boys and girls (χ^2 (df = 1) = 2.13, p = .17), and across classes (χ^2 (df = 16) = 19.88, p = .23).

3.5. Analysis strategy

3.5.1. Generalizability in item response theory (GIRT)

To study the reliability of student ratings we applied the generalizability theory in item response theory (GIRT) framework (Choi, 2012). GIRT is a combination of the item response theory (IRT) measurement model and the ANOVA model (Glas, 2012). This type of combination is also described in Briggs and Wilson (2007); Choi (2012); De Boeck (2008); Doran, Bates, Blies, and Dowling, 2007, and Fox and Glas (2001). However, only the work by Briggs and Wilson (2007); Glas (2012) and Choi (2012) explicitly link the ANOVA model with generalizability theory.

GIRT was preferred over the traditional generalizability theory, because the "My Teacher" questionnaire has been constructed and validated using Item Response Theory (IRT). Traditionally, generalizability theory is interested in the reliability of the test or form and not the items (Brennan, 2010). In generalizability theory, items typically are approached as random parallel measures, such that item scores fluctuate randomly around the parameters of the latent construct



	SLC	ECM	CSE	ATM	TLS	DII
Teacher A	✓	✗	✗	✗	✗	✗
Teacher B	✓	✓	✗	✗	✗	✗
Teacher C	✓	✓	✓	✗	✗	✗
Teacher D	✓	✓	✓	✓	✗	✗
Teacher E	✓	✓	✓	✓	✓	✗
Teacher F	✓	✓	✓	✓	✓	✓

Fig. 1. The Rasch cumulative one-dimensional scale of teaching practices. Teaching practices associated with safe learning climate (SLC) are found to generally precede teaching practices related to efficient classroom management (ECM) (i.e. no teacher is skilled in ECM without also being skilled in SLC).

(Brennan, 2010). In the Rasch (1960) model, item scores are approached as strictly parallel measures that have fixed parameters which can be estimated independently of the person parameter: teachers' teaching skill (Bond & Fox, 2007; Glas, 2012). Briggs and Wilson's (2007) GIRT model combines the one item parameter Rasch model with generalizability theory. GIRT makes no changes to the first g-study phase. Like the traditional approach to generalizability theory, a complete random effects model is estimated first to explore the item score variance decomposition. However, in the second d-study phase, the reliability is estimated for fixed item parameters and the focus is on the variance decomposition of the person parameters estimated within the structural ANOVA model.

3.5.2. Generalizability study (g-study)

In the g-study the total observed variance is decomposed in several components. These variance components are routinely referred to as "facets". For example, the facet teacher refers to the variation between teachers. Facets have levels. The term levels refers to the number of unique observations within each facet. For example, the facet teacher has levels equal to the number of teachers. The goal of the g-study is to get an impression how much variance facets contribute relative to the total variance. During the g-study phase, all facets are estimated as random effects, irrespective of whether facets are considered random or fixed in the subsequent d-study (Brennan, 2001; Shavelson & Webb, 1991). The g-study design is: students (s) crossed with teachers (t) which both are nested in classes (c) and crossed with items (i). This is technically abbreviated as: [(s × t/ct) : c] × i. In this notation, the ":" should be read as "nested in", the "/" reflects that two facets are confounded, the "×" should be read as "crossed with". The design distinguishes eight different facets.

Facets are estimated using the R package lme4 (Bates, Maechler, Bolker, & Walker, 2014). Descriptions of how to formulate and estimate multi-facet Rasch models using lme4 are available in De Boeck et al. (2011) and Doran et al. (2007). The 95% confidence interval (95% CI) of the variance components were estimated using the R package "arm" (Gelman et al., 2016). The R script used is appended in Appendix A.

3.5.3. Decision study's (d-study's)

A d-study examines the increase in the reliability that is achieved by adding more levels to a facet. The estimations make use of the variance estimates resulting from the g-study. Broadly, the d-study equation involves the ratio of the between teacher variance divided by the between teacher variance plus within teacher variance. The accuracy of the resulting reliability coefficient is dependent on how adequate the between teacher and within teacher variance(s) are estimated.

This study applies a combination of generalizability theory and IRT, referred to as GIRT. Typical to GIRT is that it considers the item facet to be fixed in the d-study (Briggs & Wilson, 2007; Choi, 2012; Glas, 2012). This implies that we approach items as having fixed parameters, such that each item measures a unique and fixed level of the construct. Because the item facet is fixed it is not included as error variance in the reliability equation. However, interactions between the item facet and other facets indicate random deviations from the unique and fixed ordering in levels (De Boeck et al., 2011). In this study these interactions are considered as error variance (see Eqs. (1) and (2)).

Two d-studies are performed. One concerning the nested-one-class-one-teacher design and one concerning the one-class-multiple-teacher design. The first d-study design is: (s : t/c) × I)). The capital letter "I" signifies that the item facet is considered fixed. The Venn diagram in Fig. 2 visualizes the variance decomposition of the design. It shows that in the nested design the between-teacher differences are described by the facet teacher (c, t, ct), which confounds the variances due to class (c), teacher (t) and class-teacher interactions (ct). Confounds indicate that variances of two or more different facets cannot be distinguished and are estimated by the one single facet. Therefore, this single facet has no unique interpretation. Its variance is the sum of the variance due to differences between classes, teachers, and teacher-class interactions (Brennan, 2001).

Note that in the Venn diagrams, the "," (instead of the "/") is used to identify confounding facets. The within teacher variance is described by two facets, namely the student facet (which in the nested design is the sum of: $\sigma_s + \sigma_{ts}$) and teacher-item interaction facet (which in the nested design is the sum of: $\sigma_{tI} + \sigma_{cI}$) plus the general error term: $(\pi^2 / 3) / (n_s n_I)$. In logistic models, effects are expressed relative to the standard logistic variance ($\pi^2 / 3$), which can be interpreted as an error (ε) (Choi, 2012; De Boeck, 2008). Choi (2012) proposes to average ε over the number of observations which she proposes is consistent with regular practice in traditional generalizability theory (e.g., Brennan, 2001).

The equation as based on generalizability theory (Brennan, 2001; Choi, 2012) is as follows:

$$\phi = \frac{\sigma^2_{c} + \sigma^2_{t,\,ct}}{\sigma^2_{c} + \sigma^2_{t,\,ct} + \frac{\sigma^2_{s} + \sigma^2_{ts}}{n_s} + \frac{\sigma^2_{tI} + \sigma^2_{cI}}{n_I} + \left(\frac{\pi^2}{3}\right)\Big/_{n_s n_I}} \tag{1}$$

The subscripts of the variances in the equation refer to the facets in the Venn diagrams (see Figs. 2 and 3) and estimated in the g-study. The second d-study involves the half-block one-class-multiple-teacher design. The Venn diagram in Fig. 3 shows that the half-block design further refines the between-teacher variance, because it splits the previously confounded facet c,t, tc into two facets, namely the facets teacher (t, ct) and class (c). In addition, the design also further refines the within-teacher student variance by splitting the student facet (s, ts) into two facets, namely the facets student (s) and teacher-student interaction (ts). Also, it further decomposes the teacher-item interaction facet (tI) into two facets, namely the teacher-item interaction (tI) and the class-item interaction (cI). This might be presumed to enhance accuracy of the reliability estimation.

The equation as based on generalizability theory (Brennan, 2001; Choi, 2012) is as follows:

$$\phi = \frac{\sigma^2_{t,\,ct}}{\sigma^2_{t,\,ct} + \frac{\sigma^2_{c}}{n_c} + \frac{\sigma^2_{s}}{n_s} + \frac{\sigma^2_{ts}}{n_s} + \frac{\sigma^2_{tI}}{n_I} + \frac{\sigma^2_{cI}}{n_c n_I} + \frac{\sigma^2_{sI}}{n_s n_I} + \left(\frac{\pi^2}{3}\right)\Big/_{n_c n_s n_I}} \tag{2}$$

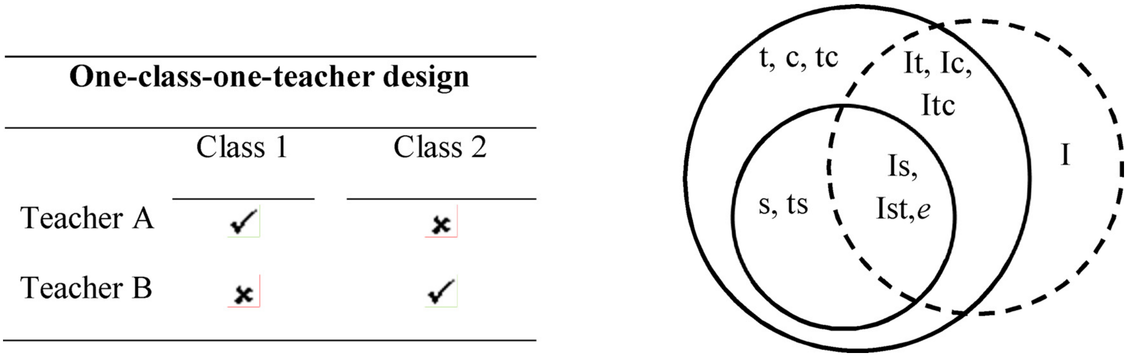
To explore how reliability changes as a function of the number of students, the parameter n_s in Eqs. (1) and (2) is varied. For each d-study design, three estimations were obtained. One using the variance components of the g-study, one using the upper 95% CI boundary of the teacher variance (keeping all other facets constant) and one using the lower 95% CI boundary of the teacher variance (keeping all other facets constant).

4. Results

The results are reported in two steps. First, the analysis concentrates on the g-study results for the one-class-multiple-teacher design. Second, the reliability of each design is examined using the d-study.

4.1. g-study

Table 3 presents the results of the g-study. The variance attributable to class and teacher variance is approximately 14%. This is somewhat lower than the 20% typically reported by studies applying the nested design (e.g., Feistauer & Richter, 2016; Marsh & Roche, 1997), but 20% is within the 95% confidence boundary [4%, 25%]. An explanation of the somewhat lower estimate of the teacher variance could be the difference in context. Perhaps that secondary school students differentiate less extremely between teachers compared to high school students or perhaps that the population of secondary education teachers shows less variation in teaching skill compared to the population of higher education teachers.



One-class-one-teacher design

	Class 1	Class 2
Teacher A	✓	✗
Teacher B	✗	✓

Fig. 2. Representation of the one-class-one-teacher design using an ANOVA design where check marks indicate that the teacher is rated by the class (left) and a Venn diagram (right). If one circle includes more than one facet, the facets are confounded.

4.2. d-studies

The d-study describes the reliability of student ratings given that the above variance decomposition (g-study) is true (see Fig. 4). The dashed line is related to the nested design and the solid line is connected to one-class-multiple-teacher design.

Fig. 4 shows that the expected reliability of the assessment differs between the two designs and is lower for the more complex one-class-multiple-teacher design. This confirms the claim of Morley (2012) that the nested design may overestimate the reliability. In the nested design, reliability is estimated to exceed 0.90 if the class aggregate is based on more than 23 students (95% CI ranges between 13 and 38 students). In the crossed one-class-multiple-teacher design, the reliability of the class aggregate exceeds the 0.90 criterion, when more than 38 students are included (95% CI ranges between 22 to 65 students).

To achieve a modest level of reliability (i.e., $\phi \geq 0.70$) required for teacher formative assessment, we see a similar shift. Based on the nested one-class-one-teacher design, an aggregate based on only as much as six students is expected to reach the reliability level of 0.70 (95% CI ranges between 4 to 8 students). Based on the one-class-multiple-teacher design, a minimum of eight students is required to reach the 0.70 criterion (95% CI ranges between 5 to 11 students). However, virtually all classes count more than eight students in practice. Hence, this finding has minor implications.

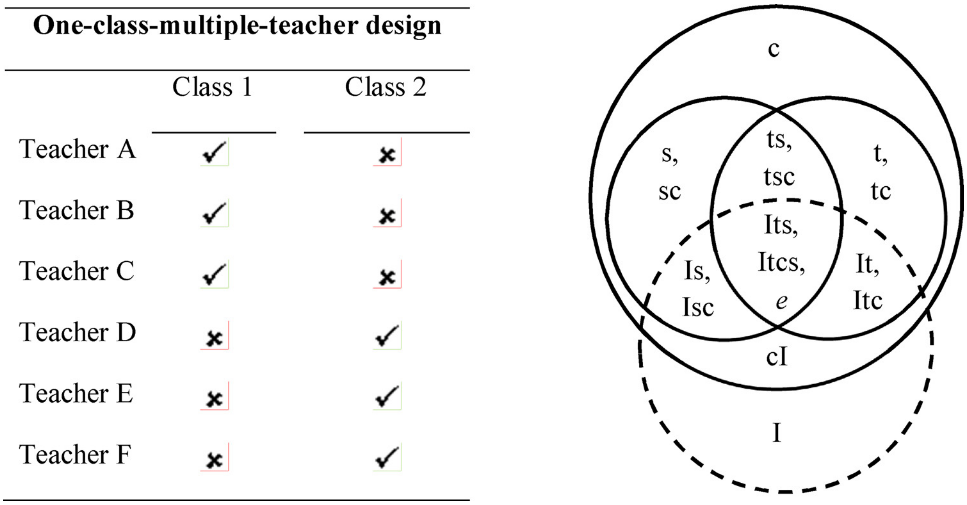
Table 3
Results of the GIRT G-study using the one-class-multiple-teacher design.

			95% CI	
Facet	σ^2	%	lower (%)	upper (%)
class (c)	0.21	3	2	8
teacher (t, tc)	0.66	11	8	16
student (s)	0.74	12	11	14
teacher × student (ts)	1.06	18	17	19
item (i)	1.91	32	22	53
item × class (ic)	0.18	3	3	3
item × teacher (it, itc)	0.53	9	8	9
item × student (is)	0.66	11	11	11

4.3. Accuracy of reliability estimation

We presented two different estimations of reliability and it is suggested that the second one-class-multiple teacher design (Fig. 4 solid line) provides the more accurate estimation. To verify this claim, we might use the observed correlation between any two randomly chosen students, and compare this with each of the two model predicted correlations. In the Fig. 4, the point "1″ on the x-axis might be interpreted as the model expected correlation of ratings given by two randomly selected students. The nested one-class-one-teacher design predicts an observed correlation of 0.33. The one-class-multiple-teacher design predicts an observed correlation of 0.25.

The observed correlation in our dataset is 0.26. This correlation corroborates Marsh's (2007) statement that the correlation between any two single students is typically in the 0.20′s. The overlap between the observed and predicted correlation provides further evidence that the one-class-multiple-teacher design provides a more accurate estimation



One-class-multiple-teacher design

	Class 1	Class 2
Teacher A	✓	✗
Teacher B	✓	✗
Teacher C	✓	✗
Teacher D	✗	✓
Teacher E	✗	✓
Teacher F	✗	✓

Fig. 3. Representation of the one-class-multiple-teacher design using an ANOVA design where check marks indicate the teachers rated by the class (left) and a Venn diagram (right). Again, if one circle includes more than one facet, the facets are confounded.

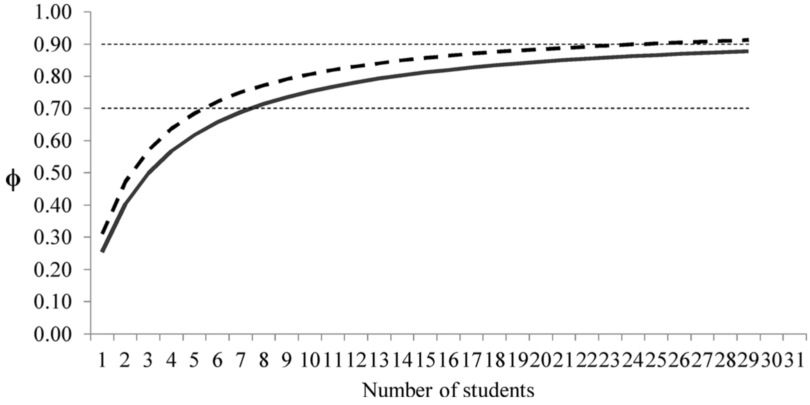


Fig. 4. The predicted reliability (ϕ) of the nested one-class-one-teacher (dashed line) and more complex one-class-multiple-teacher (solid line) design.

of the reliability of student ratings. That the predicted and observed correlation almost completely overlap should not be interpreted as suggesting the here applied design is highly accurate, because the steepness of the increase may still change if adding more complexity to the design.

5. Conclusions

The first research question concerned whether previous findings of the reliability of student ratings in higher education can be generalized to the context of secondary education. Consistent with our first hypothesis, we found that the results are similar to previous findings in higher education. Marsh (2007) summarizes previous findings by suggesting that at least 25 students are required to attain the reliability level of 0.90. Using the same nested design with a sample of secondary education students, we estimated that 23 students are required. The difference of 2 students is within the 95% confidence interval. The rapid increase of reliability observed if adding students also corroborates the claim by Kane et al. (1976) that reliability of student ratings is noticeably dependent on the number of students.

The second research question concerned whether estimations of reliability based on the one-class-one-teacher nested design are validated by a more complex design. Consistent with Morley's (2012) claim, we found that the nested one-class-one-teacher design tends to overestimate reliability of student ratings, and, hence, to underestimate the number of ratings required by different students to achieve modest (0.70) or high (0.90) reliability. The improved estimations suggest that on average ratings by eight different students are required to achieve modest reliability needed for formative assessment and ratings by 38 different students are required to achieve high reliability needed for high-stake decisions.

From the perspective of validation, the results present uncertainty about the reliability with which student ratings measure teaching skill and, thereby, also about the required number of student ratings. The reliability coefficients provided by the nested design could not be validated by the more complex design. Because the data gathering procedure does not allow for further cross-validation by using an even more complex completely crossed design, it remains an open question whether the here presented improved estimations based on the half-block are accurate.

5.1. *Why does the more complex half-block design leads to lower estimates?*

The complex design leads to lower estimates of reliability. An intuitive question might be to ask why? Variables associated to this decrease should be class characteristic. The observed decrease can not be explained by previously proposed variables, such as teacher agreeableness, teacher grading leniency, or subject matter taught (Kulik, 2001). This is, because if these variables bias student ratings they add variance to the teacher facet and not to the class facet. Variables which may explain the decrease need to be typical to the class and plausibly make that the class rates not specifically one, but *all* their teachers higher or lower compared to another class. Students age is a potential variable, but it seems negligible given the comparable results to higher education students. Another potential variable is the level of education (i.e. pre-university, higher vocational and vocational education). Though speculative, the reasoning might be that pre-university students are easier to teach (due to higher motivation and concentration) compared to students in lower vocational education due to which classes in pre-university might be more homogenous in their ratings compared to classes of students in vocational education.

5.2. *Interpretation of the index of dependability*

Generalizability theory offers two operationalizations of reliability (Brennan, 2001, 2010): (1) the generalizability coefficient provides an estimate of the consistency of the ordering between teachers, and (2) the index of dependability provides an estimate of the consistency in absolute scores. This interpretation is consistent with the most recent literature (e.g., Fan & Sun, 2014; Webb, Shavelson, & Steedle, 2012; Wiley et al., 2013). Although the here applied interpretation of the index of dependability (ϕ) has been commonly used, one may argue that the interpretation is not clear-cut. In their introduction to generalizability theory, Wiley et al. (2013) mention the generalizability coefficient and the index of dependability and discuss an extended procedure to estimate the consistency of absolute scores. This extended procedure, referred to as ϕ_λ, adds a loss function to the general formula of ϕ. The two formulas may lead to different estimates (N. M. Webb, personal communication, May 4, 2018). Given the potential implication of absolute score interpretations, future research should attend to the exact interpretation of the index of dependability in more detail.

5.3. *Implications for teacher evaluation*

The results of this study illustrate that the reliability of student ratings is not definite. It is much discussed and practiced, but still receive little attention in the literature. Although most past generalizability studies report similar reliability levels (e.g., Feistauer & Richter, 2016; Gillmore et al., 1978; Kane et al., 1976), they also applied the same research design and thereby made the same assumptions. Although the here reported reliability levels are more accurate compared to those previously reported, it is not implausible that future research, with improved and more representative design, will falsify the here reported reliability of student ratings. When making decisions based on student ratings of teaching skills, practitioners should consider the limitations of current evidence on the reliability of student ratings.

An important finding is that according to our estimates one class of students (n = 25) is insufficient to reach a reliability level of 0.90.

According to our results, ratings by 38 students are required. We note that the number of 38 is an average covering substantial individual differences. In this sample, a reliability of 0.90 was reached for 95% of the teachers if ratings by 65 students were gathered. Practitioners are cautioned to make high-stake decisions solely based on student ratings, especially if the number of ratings is lower than 38.

For formative assessment purposes, the improved estimations suggest that ratings by approximately eight students are required to obtain sufficiently reliable (≥ 0.70) aggregated scores. Also, in this sample, a random selection of ratings by approximately 11 students results in a less than 5% chance that reliability is lower than 0.70. These cut-off numbers are well below the average number of students in (Dutch) regular classes. This finding shows the potential of student rating for formative assessment purposes.

It is stressed that the reported reliability levels are to some extend dependent on the number of items, instruments used, and the (Dutch) educational context in general. Thus, though the reported reliability levels provide useful indications, we caution generalizing results to other contexts.

5.4. Implications for research

The study indicates that the nested design may leads to an overestimation of the teacher variance. This also has implications for using student ratings as a predictor or an outcome variable in research settings. Martínez (2012) discusses the consequences of omitting the class facet when studying student achievement data. Although Martinez's study concerns another dependent variable (reading achievement), his findings provide some suggestions as regards how omitting important facets might distort research findings. Most importantly, he reports that neglecting important facets might give misleading results which only surface if the facet is omitted.

The here applied one-class-multiple-teacher (half block) design delivers evidence to improve on the nested design. Nevertheless, room for further improvement is suggested. An inspection of the results in Table 3 will show that the here applied multiple-class-one-teacher design does not adequately address the teacher × class (tc) interaction facet. This facet currently is confounded with the teacher facet (t) still leading to an overestimation in the differences between teachers. To address the effect of this bias appropriately, future research should sample student ratings using a design which includes multiple classes rating the same group of teachers. This may either lead to further improvement on the estimation of the number of students required to achieve modest and high levels of reliability, or validate the in this study reported reliability levels as being accurate.

Finally, the present study findings greatly overlap with the previous results reported by Gillmore et al. (1978) and Kane et al. (1976), by applying "My Teacher" questionnaire containing dichotomous response categories. Furthermore, additional analyses using a polytomous version of the "My Teacher" questionnaire revealed negligible differences in terms of reliability estimates with the dichotomous version reported in this study[2]. Reasons for this finding are not straightforward. Although polytomous categories allow for more variation in responses, and reliability estimation depends on the variance components, our findings seem to suggest that response categories do not matter much for reliability estimations. This is especially true as far as the "My Teacher" questionnaire is concerned.

A possible explanation for this is that, the reliability estimation is not dependent to the absolute size of the total variance, but on the relative percentage of variation attributable to each component. The results might hint that the total variance is decomposed equally in both polytomous and dichotomous response categories. Another interpretation of this result is that polytomous and dichotomous item responses might show a similar level of dependency on the latent variable of teaching skill. This could be viewed as providing support for the use of dichotomous response categories.

5.5. Limitations

An assumption underlying the generalizability theory is that the sample variances reported in the g-study and upon which the d-study is based are accurate estimates of the population variances. The study design sampled three to four teachers per class and estimation of the class variance is thus dependent on three-four measurement points. Whether this number is sufficient to provide an accurate estimation of the class variance might be subject for future research.

Another limitation involves the possibility of confounds. One potential confounding factor concerns the subject taught. In the current design, every teacher is rated by merely one class confounding the subject and teachers. Thus, the between-teacher variation as rated by the class might also be interpreted as the between-subject variation. Because the study predominantly sampled the same four subjects, generalization of findings beyond these subjects should be done cautiously.

The imbalance in sample composition with respect to gender, in which 60.5% teachers were male, is also reason for concern. This number is unrepresentative for the Dutch teacher workforce, which is dominated by female. There is only few evidence suggesting that the overrepresentation of male teachers might have resulted in biased estimates. Replication studies with a more gender representative sample may further strengthen the conclusions.

Finally, one may argue that Choi (2012) GIRT method based on the logistic link has not been rigorously examined. Her analyses as well as our own analyses suggest that results are comparable (though not exactly identical) to results reported by studies using regular generalizability theory methods. However, much remains unknown about this method. Future research might benefit from taking a probit link into consideration as a complementary to the logistic link used in this study.

[2] Results can be requested by the first author

Appendix A. R script used in the g-study

```
# G-study student ratings Table and Figure 1.
library(lme4)
library(arm)
Data.Gstudy = read.table("file location”,


                         header = T, sep = ",")
Data.Gstudy$Itemnumber.f = factor(Data.Gstudy$Itemnumber)
Data.Gstudy$Case.f = factor(Data.Gstudy$ï..Case)
Data.Gstudy$Teacher.f = factor(Data.Gstudy$Teacher)
Data.Gstudy$Class.f = factor(Data.Gstudy$Class)
Data.Gstudy$Student.f = factor(Data.Gstudy$Student)
Data.Gstudy$Domain.f = factor(Data.Gstudy$Domain)

# Partially nested g-study (half block)
GIRT.halfblock = glmer(Response ~ 1 + (1 | Itemnumber.f) +
(1 | Teacher.f) + (1 | Student.f) + (1 | Case.f) + (1 | Class.f) +
(1 | Itemnumber.f : Student.f) + (1 | Itemnumber.f : Teacher.f) +
(1 | Itemnumber.f : Class.f), data=Data.Gstudy, binomial)

summary(GIRT.halfblock)

#95% CI halfblock
GIRT.halfblock.Var.CL = VarCorr(GIRT.halfblock)$Class.f[1]
GIRT.halfblock.N.CL = nrow(ranef(GIRT.halfblock)$Class.f)
GIRT.halfblock.CI.CL = (GIRT.halfblock.N.CL - 1) * GIRT.halfblock.Var.CL /
qchisq(c(0.975, 0.025), df = GIRT.halfblock.N.CL - 1)
GIRT.halfblock.Var.Tea = VarCorr(GIRT.halfblock)$Teacher.f[1]
GIRT.halfblock.N.Tea = nrow(ranef(GIRT.halfblock)$Teacher.f)
GIRT.halfblock.CI.Tea = (GIRT.halfblock.N.Tea - 1) * GIRT.halfblock.Var.Tea
/ qchisq(c(0.975, 0.025), df = GIRT.halfblock.N.Tea - 1)
GIRT.halfblock.Var.St = VarCorr(GIRT.halfblock)$Student.f[1]
GIRT.halfblock.N.St = nrow(ranef(GIRT.halfblock)$Student.f)
GIRT.halfblock.CI.St = (GIRT.halfblock.N.St - 1) * GIRT.halfblock.Var.St /
qchisq(c(0.975, 0.025), df = GIRT.halfblock.N.St - 1)
GIRT.halfblock.Var.StuTea = VarCorr(GIRT.halfblock)$case.f[1]
GIRT.halfblock.N.StuTea = nrow(ranef(GIRT.halfblock)$case.f)
GIRT.halfblock.CI.StuTea = (GIRT.halfblock.N.StuTea - 1) *
GIRT.halfblock.Var.StuTea / qchisq(c(0.975, 0.025), df =
GIRT.halfblock.N.StuTea - 1)
GIRT.halfblock.Var.IT = VarCorr(GIRT.halfblock)$Itemnumber.f[1]
GIRT.halfblock.N.IT = nrow(ranef(GIRT.halfblock)$Itemnumber.f)
GIRT.halfblock.CI.IT = (GIRT.halfblock.N.IT - 1) * GIRT.halfblock.Var.IT /
qchisq(c(0.975, 0.025), df = GIRT.halfblock.N.IT - 1)
GIRT.halfblock.Var.ITCL = VarCorr(GIRT.halfblock)$`Itemnumber.f:Class.f`[1]
GIRT.halfblock.N.ITCL = nrow(ranef(GIRT.halfblock)$`Itemnumber.f:Class.f`)
GIRT.halfblock.CI.ITCL = (GIRT.halfblock.N.ITCL - 1) *
GIRT.halfblock.Var.ITCL / qchisq(c(0.975, 0.025), df =
GIRT.halfblock.N.ITCL - 1)
GIRT.halfblock.Var.ITTE =
VarCorr(GIRT.halfblock)$`Itemnumber.f:Teacher.f`[1]
GIRT.halfblock.N.ITTE =
nrow(ranef(GIRT.halfblock)$`Itemnumber.f:Teacher.f`)
GIRT.halfblock.CI.ITTE = (GIRT.halfblock.N.ITTE - 1) *
GIRT.halfblock.Var.ITTE / qchisq(c(0.975, 0.025), df =
GIRT.halfblock.N.ITTE - 1)
GIRT.halfblock.Var.ITST =
VarCorr(GIRT.halfblock)$`Itemnumber.f:Student.f`[1]
GIRT.halfblock.N.ITST =
nrow(ranef(GIRT.halfblock)$`Itemnumber.f:Student.f`)
GIRT.halfblock.CI.ITST = (GIRT.halfblock.N.ITST - 1) *
GIRT.halfblock.Var.ITST / qchisq(c(0.975, 0.025), df =
GIRT.halfblock.N.ITST - 1)
```

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using S4 classes. R package version 1*. 1–7. http://CRAN.R-project.org/package=lme4.

Benton, S. L., & Cashin, W. E. (2012). *Student ratings of teaching: A summary of the research and literature. (IDEA paper No. 50)*. Retrieved March 3, 2015, fromhttp://www.ntid.rit.edu/sites/default/files/academic_affairs/Sumry%20of%20Res%20%2350%20Benton%202012.pdf.

Bill & Melinda Gates Foundation (2012). *Asking students about teaching: Student perception surveys and their implementation.* Retrieved March 3 fromhttp://www.metproject.org/downloads/Asking_Students_Practitioner_Brief.pdf.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences.* London: Lawrence Erlbaum Associates.

Brennan, R. L. (2001). *Generalizability theory: Statistics for social science and public policy.* New York: Springer-Verlag.

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*(1), 1–21. http://dx.doi.org/10.1080/08957347.2011.532417.

Briggs, D. C., & Wilson, M. (2007). Generalizability in item response modelling. *Journal of Educational Measurement, 44*, 131–155.

Choi, J. (2012). *Advances in combining generalizability theory and item response theory. Doctoral dissertation.* Berkeley: University of California.

Cronbach, L. J., Gleser, C. G., Rajaratnam, N., & Nanda, H. (1972). *The dependability of behavioral measurements.* New York, USA: Wiley.

De Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*(4), 533–559.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Abe, H., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*, 1–25.

Doran, H., Bates, D., Blies, P., & Dowling, M. (2007). Estimating the multilevel Rasch model with the lme4 package. *Journal of Statistical Software, 20*, 1–18.

Fan, X., & Sun, S. (2014). Generalizability theory as a unifying framework of measurement reliability in adolescent research. *The Journal of Early Adolescence, 34*(1), 38–65.

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction, 29*, 1–9. http://dx.doi.org/10.1016/j.learninstruc.2013.07.001.

Feistauer, D., & Richter, T. (2016). How reliable are students' evaluations of teaching quality? A variance components approach. *Assessment & Evaluation in Higher Education,* 1–17.

Fox, J. P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*, 271–288.

Gelman, A., Su, Y.-S., Yajima, M., Hill, J., Pittau, M. G., Kerman, J., ... Dorie, V. (2016). *Arm: Data analysis using regression and multilevel/ hierarchical models. R package version 1.9-3.* 2016https://cran.r-project.org/web/packages/arm/arm.pdf.

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings of instruction: Estimation of teacher and course components. *Journal of Educational Measurement, 15*(1), 1–13.

Glas, C. A. W. (2012). Generalizability theory and item response theory. In T. J. H. M. Eggen, & B. P. Veldkamp (Eds.). *Psychometrics in practice at RCEC*http://dx.doi.org/10.3990/3.9789036533744.ch1 (pp. 1–13). E-book, Adobe pdf version.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel. Research paper. MET project.* Bill & Melinda Gates Foundation.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. http://dx.doi.org/10.1111/jedm.12000.

Kane, M. T., & Brennan, R. L. (1977). The generalizability of class means. *Review of Educational Research, 47*(2), 267–292.

Kane, M. T., Gillmore, G. M., & Crooks, T. J. (1976). Student evaluations of teaching: The generalizability of class means. *Journal of Educational Measurement, 13*(3), 171–183.

Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. *New Directions for Institutional Research, 2001*(109), 9–25.

Lok, B., McNaught, C., & Young, K. (2016). Criterion-referenced and norm-referenced assessments: Compatibility and complementarity. *Assessment & Evaluation in Higher Education, 41*(3), 450–465. http://dx.doi.org/10.1080/02602938.2015.1022136.

Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *Learning Environments Research, 9*(3), 215–230.

Marsh, H. W. (1982). The use of path analysis to estimate teacher and course effects in student ratings of instructional effectiveness. *Applied Psychological Measurement, 6*(1), 47–59.

Marsh, H. D. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry, & J. C. Smart (Eds.). *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). Dordrecht, The Netherlands: Springer.

Marsh, H. W., & Hocevar, D. (1991). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and Teacher Education, 7*(4), 303–314.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187–1197.

Martínez, J. F. (2012). Consequences of omitting the classroom in multilevel models of schooling: An illustration using opportunity to learn and reading achievement. *School Effectiveness and School Improvement, 23*(3), 305–326.

Marzano, R. J., & Toth, M. D. (2013). *Teacher evaluation that makes a difference. A new model for teacher growth and student achievement.* Alexandria, Virginia: ASCD.

Maulana, R., Helms-Lorenz, M., & van de Grift, W. (2015). Development and evaluation of a questionnaire measuring pre-service teachers' teaching behaviour: A Rasch modelling approach. *School Effectiveness and School Improvement, 26*(2), 169–194. http://dx.doi.org/10.1080/09243453.2014.939198.

Morley, D. D. (2012). Claims about the reliability of student evaluations of instruction: The ecological fallacy rides again. *Studies in Educational Evaluation, 38*(1), 15–20.

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill.

Panayiotou, A., Kyriakides, L., Creemers, B. P., McMahon, L., Vanlaar, G., Pfeifer, M., ... Bren, M. (2014). Teacher behavior and student outcomes: Results of a European study. *Educational Assessment, Evaluation and Accountability, 26*(1), 73–93. http://dx.doi.org/10.1007/s11092-013-9182-x.

Peterson, K. D., Wahlquist, C., & Bone, K. (2000). Student surveys for school teacher evaluation. *Journal of Personnel Evaluation in Education, 14*(2), 135–153.

Polikoff, M. S. (2015). The stability of observational and student survey measures of teaching. *American Journal of Education, 121*(2), 183–212. http://dx.doi.org/10.1086/679390.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Nielsen & Lydiche.

Richardson, J. T. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment & Evaluation in Higher Education, 30*(4), 387–415.

Rindermann, H., & Schofield, N. (2001). Generalizability of multidimensional student ratings of university instruction across courses and teachers. *Research in Higher Education, 42*(4), 377–399.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Thousand Oaks, California: Sage Publications, Inc.

van der Lans, R. M., van de Grift, W. J. C. M., & van Veen, K. (2015). Developing a teacher evaluation instrument to provide formative feedback using student ratings of teaching acts. *Educational Measurement: Issues and Practice, 34*(3), 18–27. http://dx.doi.org/10.1111/emip.12078.

Webb, N. M., Shavelson, R. J., & Steedle, J. T. (2012). Generalizability theory in assessment contexts. In C. Secolsky, & D. B. Denison (Eds.). *Handbook on measurement, assessment, and evaluation in higher education* (pp. 132–149). London, UK: Routledge.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness. New teacher project*.

Wiley, E. W., Webb, N. M., & Shavelson, R. J. (2013). The generalizability of test scores. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.). *APA handbooks in psychology. APA handbook of testing and assessment in psychology vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (pp. 43–60). Washington, DC, US: American Psychological Association.