Studies in Educational Evaluation xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Studies in Educational Evaluation

journal homepage: www.elsevier.com/stueduc

# Assessing the validity and reliability of a quick scan for student's evaluation of teaching. Results from confirmatory factor analysis and G Theory

Pieter Spooren *, Dimitri Mortelmans, Wim Christiaens

*Department of Social and Political Sciences, University of Antwerp, Antwerp, Belgium*

ARTICLE INFO

*Article history:*
Received 3 December 2013
Received in revised form 26 February 2014
Accepted 4 March 2014

*Keywords:*
Student evaluation of teaching
Teacher evaluation
Validity
Higher education
G Theory
Confirmatory factor analysis

ABSTRACT

Using confirmatory factor analysis and G Theory analysis, this article explores the reliability and the validity of a short version of the SET37 questionnaire for students' evaluation of teaching (SET). The results show that this instrument can be used as a valuable diagnostic instrument for gathering student feedback in internal practices and procedures aimed at both monitoring and improving the quality of instruction in higher education.

© 2014 Elsevier Ltd. All rights reserved.

## Introduction

Nowadays, student evaluation of teaching (SET) is used as a measure of teaching performance in almost every institution for higher education throughout the world (Zabaleta, 2007). Universities and university colleges have developed more or less complex procedures and instruments to collect, analyze and interpret these data as the dominant and sometimes sole indicator of teaching quality (Onwuegbuzie et al., 2007). This widespread use has much to do with their (apparent) ease of collecting the data and presenting and interpreting the results. When using student surveys, which are subject of most published SET research, SET-practice mainly comes down to this:

> "At the end of a semester (or at the start of the next semester), students evaluate all instructors in every course offered during that semester. They use a general survey instrument that is applicable to as many types of courses as possible and contains questions concerning teaching skills, organization of the course, workload, study materials et cetera. These questions are answered by means of a Likert-type scale ranging between 'not good at all' ('totally disagree') to 'very good' ('totally agree'). Most instruments have also additional space for students' written comments. To preserve students' anonymity, the surveys are conducted (during class time or using web-based surveys) by administrators, usually in the absence of the teacher. The results are presented in a SET-report that usually contains both a quantitative overview of the responses to the Likert-scales (mean scores, standard deviances, histograms) and all written comments. This report is shared with the teacher (for the improvement of teaching in that particular course) and the institutional board (for summative decision-making)." (Spooren, 2012, p. 4)

This double use (i.e. for both improvement and evaluation) makes the use of SET very delicate (Penny, 2003). On the one hand, many teachers are convinced of the usefulness of SET as an instrument for feedback on their teaching (Richardson, 2005). SET results help them to improve the quality of their teaching as it provides them with useful insights in the strengths and weaknesses of their teaching practice, based on student opinions. On the other hand, it is argued that nowadays the principal purpose of SET lies in its use as a measure for quality monitoring, administrative policy-making and mapping whether or not teachers reach a certain required standard in their teaching practice (Chen & Hoshower, 2003; Douglas & Douglas, 2006; Penny & Coe, 2004). This justification for using SET in staff appraisals is related to an increasing focus on internal quality assurance and performance management in universities, which have become subject to the

* Corresponding author at: Department of Social and Political Sciences, Sint Jacobstraat 2, B-2000 Antwerp, Belgium. Tel.: +32 3 265 53 60; fax: +32 3 265 57 93.
*E-mail address:* pieter.spooren@ua.ac.be (P. Spooren).

http://dx.doi.org/10.1016/j.stueduc.2014.03.001
0191-491X/© 2014 Elsevier Ltd. All rights reserved.

Please cite this article in press as: P. Spooren, et al.. Assessing the validity and reliability of a quick scan for student's evaluation of teaching. Results from confirmatory factor analysis and G Theory. *Studies in Educational Evaluation* (2014), http://dx.doi.org/10.1016/j.stueduc.2014.03.001

demands of consumer satisfaction (Blackmore, 2009). Teacher performance and the quality of teaching could be defined as the extent to which student expectations are met, thus equating student "opinions" with "teaching quality".

For this reason, many faculty members have been questioning the validity and reliability of SET results for many years (Ory, 2001). In general, their concerns include (a) the differences between the ways in which students and teachers perceive effective teaching, (b) the relationships between SET scores and factors that are unrelated to "good teaching" (Centra, 2003; Marsh, 2007), (c) SET procedures and practices (the contents of SET reports, the depersonalization of the individual relationship between teachers and their students due to the standardized questionnaires and respondents' anonymity, the competency of SET administrators, the low response rates, etc.), and (d) the psychometric value of the SET instruments.

Regarding the latter, a common understanding and a conceptual framework concerning the concept of effective teaching, upon which SET-instruments could be grounded, still does not exist (Onwuegbuzie, Daniel, & Collins, 2009; Penny, 2003). Moreover, there is no consensus in the SET literature on the number and the nature of dimensions concerning effective teaching that should be captured in SET instruments (Jackson et al., 1999). As a result, SET instruments vary greatly in both content and construction, due to the characteristics and desires of particular institutions. Besides, many institutions make use of ad hoc instruments that were not tested at all (Richardson, 2005).

Still, several well-designed instruments for measuring students' observations concerning the quality of (teaching in) a course are available. Examples are the Students' Evaluations of Educational Quality or SEEQ (Marsh, 1982; Marsh et al., 2009), the Course Experience Questionnaire or CEQ (Ramsden, 1991; Wilson, Lizzio, & Ramsden, 1997), the Student Course Experience Questionnaire or SCEQ (Ginns, Prosser, & Barrie, 2007), the SET37 (Mortelmans & Spooren, 2009; Spooren, Mortelmans, & Denekens, 2007), the Exemplary Teacher Course Questionnaire or ETCQ (Kember & Leung, 2008), and the Teaching Behavior Checklist (Keeley, Smith, & Buskist, 2006; Keeley, Furr, & Buskist, 2010). These instruments have in common that they are grounded upon educational theory and that they were tested and re-tested extensively on multiple aspects concerning their validity and reliability. Moreover, the various dimensions concerning effective teaching are measured by means of Likert scales in which sets of items measure several dimensions of teaching quality (which are seen as latent constructs). These scales allow both rigorous tests of the instrument's psychometric properties and a straightforward quality check (e.g., by calculating alpha statistics) for each dimension contained in a SET report. Table 1 contains an overview of the number and the nature of the dimensions that are measured in the aforementioned instruments.

Still, each of these instruments contains a high number of items since its developers wanted to capture as much dimensions of teaching as possible (to provide detailed feedback to the teachers being evaluated). Such instruments may overburden students, who are among a heavily surveyed group, especially when they are invited into a high number of course evaluations in a short-time period (Spooren & Van Loon, 2012). As a consequence, SET could suffer from non-response and/or biased results (i.e. acquiescence, response patterns, being too critical). Institutions, educational policy-makers, and teachers therefore need instruments that are much shorter and can be used as short-form screening instruments rather than as elaborate tests of a teacher's competences regarding each dimension of teaching in a particular course. Such instruments could provide a quick evaluation of a course, and may be followed by an evaluation with a larger and more powerful instrument if necessary. Still, it is very important to be aware of the reliability and the validity of these 'quick scans' since one might suppose that results from such instruments will be used for both formative and summative purposes.

## Objectives

Although recent research in the field has shown that single-item ratings of instructional quality are highly reliable (Ginns & Barrie, 2004; Wanous & Hudy, 2001), we believe that SET by means of one question (e.g., 'Overall I was satisfied with the quality of this course') may not be very helpful for both monitoring teaching quality and/or the improvement of teaching as this practice assumes that quality of instruction can be observed unequivocally. Besides, it has been shown that single-item ratings are, in many cases, less accurate and less valid than multi-item scales (Marsh, 1987). A quick scan instrument should therefore at least cover some important dimensions of teaching that grant a first indication of students' perceptions regarding these topics, rather than revealing a class average score on one general item that does not offer any insight in those aspects of teaching that may need attention.

The recent observation that SET scores on several dimensions of teaching could be captured by a second-order factor that represents a global construct (i.e., a general instructional skill) (Apodaca & Grad, 2005, Burdsal & Harrison, 2008, Ginns & Barrie, 2009; Spooren and Mortelmans, 2009) could be used as a starting point for constructing a quick scan instrument. After all, items that have high loadings on these first-order dimensions of teaching could be estimated as direct measures of such a second-order global factor. In the present study, we discuss the construction and validation procedure of such an instrument (which consists of 9 items) that was derived from the SET37 questionnaire for student's evaluation of teaching (Mortelmans & Spooren, 2009) which is used for SET at the authors' institution.

The basic assumptions for the questionnaire were that it should (a) be reliable and valid, (b) be short, and (c) contain questions on several important dimensions of instruction that are captured in the SET37 questionnaire. Moreover, the questionnaire should perform well with different kinds of courses and different kinds of disciplines. Therefore, the sampling procedure needs to cover as much as possible disciplines and levels of study. Besides, to allow testing the instrument on its stability, data should be collected at two or more different occasions.

## Method

### *Sample*

SET were administered in class during the fall semester of the 2012–2013 academic year at the University of Antwerp (a medium-sized Belgian university with approximately 13 000 students). Students from 6 faculties (Science, Social and political sciences, Law, Literature and Philosophy, Economics, and Pharmaceutical, biomedical and veterinary sciences) evaluated 16 courses two times (at an interval of 3–4 weeks). So students filled in the questionnaire twice, to give feedback about a course they had followed during the 2011–2012 academic year. The number of students who evaluated a course ranged from 16 to 188 (time point 1, 1139 students) and from 12 to 166 (time point 2, 941 students). The mean numbers of respondents were 71 (time point 1, SD = 51) and 59 (time point 2, SD = 39), the median number of respondents were lower (65 in time point 1, 54.5 in time point 2). Students were also asked to provide some identification details, which were used to link the questionnaires from both moments. 641 students completed the questionnaire at both time points, which allowed us

Please cite this article in press as: P. Spooren, et al.. Assessing the validity and reliability of a quick scan for student's evaluation of teaching. Results from confirmatory factor analysis and G Theory. *Studies in Educational Evaluation* (2014), http://dx.doi.org/10.1016/j.stueduc.2014.03.001

**Table 1**
Summary of dimensions in SET-instruments.

| Author | Instrument | No. of dimensions | Dimensions | No. of items |
|---|---|---|---|---|
| Marsh et al. (2009)<br>Marsh (1982) | SEEQ | 9 | Learning/value<br>Instructor enthusiasm<br>Organization/clarity<br>Group interaction<br>Individual rapport<br>Breadth<br>Exam/graded materials<br>Readings/assignments<br>Workload difficulty | 31 |
| Ramsden (1991)<br>Wilson et al. (1997) | CEQ | 5 | Good teaching<br>Clear goals and standards<br>Appropriate workload<br>Appropriate assessment<br>Emphasis on independence | 57 |
| Ginns et al. (2007) | SCEQ | 5 | Good teaching<br>Clear Goals and standards<br>Appropriate assessment<br>Appropriate workload<br>Generic skills | 23 |
| Spooren et al. (2007)<br>Mortelmans and Spooren (2009)<br>Spooren, Mortelmans, and Van Loon (2012) | SET37 | 12 | Clarity of objectives<br>Value of subject matter<br>Build-up of subject matter<br>Presentation skills<br>Harmony organization course–learning<br>Course materials<br>Course difficulty<br>Help of the teacher during the learning process<br>Authenticity of the examination(s)<br>Linking-up with foreknowledge<br>Content validity of the examination(s)<br>Formative evaluation(s) | 37 |
| Kember and Leung (2008) | ETCQ | 9 | Understanding fundamental content<br>Relevance<br>Challenging beliefs<br>Active learning<br>Teacher-student relationships<br>Motivation<br>Organization<br>Flexibility<br>Assessment | 27 |
| Keeley et al. (2006)<br>Keeley et al. (2010) | TBC | 2 | Caring and supportive<br>Professional competency and Communicational skills | 28 |

*Note.* Keeley et al. (2006) found a good fit for one-factor model to the data as well.

to execute a number of reliability and stability tests during the validation procedure (see below).

*Measures*

The present study was part of a re-validation procedure of the SET37 questionnaire. The original SET37 questionnaire consists of 12 quasi-balanced Likert scales representing 12 dimensions of teaching. The reworked version that the students filled in omitted 1 scale and slightly reworked the items of the remainder 11 scales representing 11 dimensions of teaching: Clarity of objectives, Build-up of subject matter, Presentation skills, Harmony between organization of the course and the student's learning process, Course materials, Course difficulty, Help of the teacher during the learning process, Authenticity of the examination, Linking-up with advance knowledge, Content validity of the examination, and Formative examination(s). Each of these dimensions is measured with at least three items. The reworked SET37 questionnaire also added three single-item questions to the original SET37 instrument (overall quality of the course, student learning, and the relevance of this course for the educational program). All items are scored on a six-point scale, going from 'strongly disagree' to 'strongly agree', except the two open-ended questions allowing free response at the end of the questionnaire. Students were given the same instructions by the authors before completing the course evaluations.

During this re-validation procedure,[1] factor analysis confirmed both the factor structure of the SET37 questionnaire and the existence of a second-order factor behind six of the eleven scales in the instrument (Mortelmans & Spooren, 2009; Spooren, 2010; Spooren et al., 2007), including Clarity of objectives, Build-up of subject matter, Presentation skills, Course materials, Course difficulty, and Help of the teacher during the learning process. It was assumed that this second-order factor can be considered a 'teacher professionalism factor' as it influences these factors that measure the way a teacher built up, organized and executed his/her course (Spooren and Mortelmans, 2006). If he/she managed to do this professionally, this will be rewarded by the students as 'good teaching' and thus with higher ratings on the six scales in the reworked version of the SET37 questionnaire.

[1] The results of this procedure are available on request. The rationale for the re-validation of the SET37 was to confirm the psychometric properties of the instrument (which is also used for personnel decisions at this institution) after making some minor changes (i.e. item wordings) and the replacement of one scale ('Relevance of subject matter') by a single item type question.

Please cite this article in press as: P. Spooren, et al.. Assessing the validity and reliability of a quick scan for student's evaluation of teaching. Results from confirmatory factor analysis and G Theory. *Studies in Educational Evaluation* (2014), http://dx.doi.org/10.1016/j.stueduc.2014.03.001

**Table 2**
The SET37_QS Instrument for student evaluation of teaching.

| Item No. | Item | SET37 dimension | Factor loading item on SET37 dimension |
|---|---|---|---|
| QS1 | In this course, it was clearly specified what I should learn and accomplish at the end of the course | Clarity of course objectives | .94 |
| QS2 | The various themes in this course were well geared to one another | Build-up of subject matter | .97 |
| QS3 | The lecturer explained the subject matter well | Presentation skills | .96 |
| QS4 | The study materials were useful | Course materials | .94 |
| QS5 | The teacher's expectations to what I should have learned at the end of the course were realistic and acceptable | Course difficulty | .95 |
| QS6 | The teacher helped me with questions and problems which arose during this course | Help of the teacher during the learning process | .93 |
| QS7 | Overall, I am satisfied with this course | *Single item* | – |
| QS8 | I have learned a lot during this course | *Single item* | – |
| QS9 | In my understanding, this course is relevant to my educational program | *Single item* | – |

*Note.* Translated from Dutch. Factor loadings are standardized. The instrument also contains two open-ended questions, i.e. 'Which were, in your opinion, the strengths of this course and should be retained?' and 'Which were, in your opinion, the weaknesses of this course and should be improved?'

### *Analytic strategy*

For the development of the quick scan instrument, we selected from each of the aforementioned six scales the positive worded item with the highest loading on its construct (based on the data we collected at time point 1). In addition, the three single-item type questions were included: Relevance of subject matter,[2] Students' subjective perception of learning and Overall satisfaction of the course. Table 2 contains an overview of all nine items in the short questionnaire (further: SET37_QS) and the teaching dimensions associated with each item.

To assess the psychometric properties of the proposed instrument, multiple analyses were conducted. After exploring the inter-item correlations, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) was used to gain more insight in the factor structure of the instrument and to determine several aspects of its validity and reliability.

In a second phase, we conducted a G study (Brennan, 2001) for estimating the reliability of the SET37_QS, using data from both times. The reliability of SET instruments and many other measures in educational sciences is commonly assessed by means of classical test theory, which provides a number of methods such as internal consistency, test-retest and interobserver agreement analyses. Still, these tests can only be done one at a time (to check for one source of measurement error each time) and cannot be combined to a test that provides an overall estimate of reliability (Mushquash & O'Connor, 2006). On top of that, researchers are not given any information or suggestions concerning the number of items or occasions that are needed to obtain a reliable measure (Webb, Rowley, & Shavelson, 1988). Generalizability theory, which liberalizes classical theory through an application of ANOVA methods to measurement issues (Brennan, 2001), provides a useful alternative as it combines many aspects of classical test theory into one overall estimate of reliability (disentangling both multiple sources of error and interactions between these sources). When the overall estimate, the so-called *G coefficient* is high, the obtained scores from an instrument can be generalized across the several facets that are included in the study (for instance, different occasions, respondents and/or items). Besides, researchers are, by means of a *D study*, informed about improvements on measurement reliability by, for instance, the number of items in a questionnaire, the number of respondents, and/or the number of occasions (Mushquash & O'Connor, 2006).

## Results

### *Correlational analysis*

At time point 1, the interitem correlations[3] between the items across all courses were statistically significant and generally moderate to large as they ranged between .25 and .59 (Cohen, 1992). The correlations between the 'overall' item and the other items ranged between .35 and .59, which provides a first indication of the convergent validity of the items in the questionnaire (Ginns & Barrie, 2009). At the individual course level however, interitem correlations showed greater variety (with correlations ranging between −.34 and .92). This is an interesting finding as it suggests that the instrument allows students to underline important points (strengths and/or weaknesses) concerning several aspects of teaching within a particular course, apart from other dimensions of teaching (including the 'overall' item) that are measured by the instrument.

### *Exploratory factor analysis*

Exploratory factor analysis was used to screen out possibly problematic items in the instrument (i.e. high item loadings on nonhypothesized factors). The analysis, which was based on the data collected at time point 1, used unweighted least squares estimation at the student level and extracted one factor with an eigenvalue greater than 1 (4.418). This factor includes all items from the SET37_QS (with item loadings ranging between .43 and .82) and explains 43.1% of the variance. These findings indeed suggest that the items from the SET37_QS belong to a global factor, such as general instructional skill, which explains a great amount in SET scores.

### *Confirmatory factor analysis (CFA)*

As the data in the present study have a hierarchical structure (students are nested in courses and individual observations

[2] The original SET37 questionnaire also included a scale for the dimension 'Relevance of subject-matter' which was deleted in the reworked version and replaced by a single-item type question. This dimension had high loadings on the 'teacher professionalism' factor as well (Mortelmans and Spooren, 2009). On that account, we added an item from the original scale to the quick scan instrument.

[3] Correlation matrices are available on request.

Please cite this article in press as: P. Spooren, et al.. Assessing the validity and reliability of a quick scan for student's evaluation of teaching. Results from confirmatory factor analysis and G Theory. *Studies in Educational Evaluation* (2014), http://dx.doi.org/10.1016/j.stueduc.2014.03.001

**Table 3**
Standardized factor loadings (Est.), corrected standard errors (SE), and intraclass correlations (ICC) at time point 1 (t1) and time point 2 (t2) of the SET37_QS items.

| Item | t1 | | | t2 | | |
|---|---|---|---|---|---|---|
| | Est. | SE | ICC | Est. | SE | ICC |
| QS1 | .61 | .05 | .11 | .67 | .03 | .12 |
| QS2 | .65 | .03 | .14 | .64 | .03 | .12 |
| QS3 | .67 | .04 | .16 | .65 | .03 | .14 |
| QS4 | .65 | .05 | .15 | .67 | .05 | .12 |
| QS5 | .64 | .03 | .20 | .65 | .03 | .16 |
| QS6 | .43 | .05 | .15 | .46 | .05 | .16 |
| QS7 | .82 | .03 | .17 | .68 | .04 | .11 |
| QS8 | .69 | .05 | .12 | .73 | .03 | .12 |
| QS9 | .68 | .05 | .24 | .69 | .04 | .21 |

of teaching skills thus are not completely independent) (Nasser & Hagtvet, 2006; Spooren, 2010; Wagner et al., 2013), this structure should be taken into account when analyzing the factorial structure of the data (Kreft & De Leeuw, 1998; Hox, 2010). Inspection of the intraclass correlation coefficients (Table 3) shows that a relevant component of the variance (ICC's ranged between .11 and .24) in the items of the SET37_QS was due to the course level, which indicates students' shared perceptions of instructional quality (Wagner et al., 2013).

Still, a full multilevel CFA procedure as proposed by Muthén (1994) was not possible due to the small number of units at the second level (course level, $N = 16$) and, especially, the third level (Department, $N = 6$). We however took into account this complex sample structure by running CFA models with the 'type = complex feature' in the MPlus software (Muthén & Muthén, 1998–2010), which specified the course level in the analysis and allows to apply corrected standard errors for all study variables.

CFA models were run for both times we collected data and specified one global factor. The standardized factor loadings with the corrected standard errors are shown in Table 3. All loadings are statistically significant and substantial. This suggests that all 9 items from the SET37_QS adequately reflect the same construct. The one-factor solution is also invariant over the two time points as the factor loadings remain quite stable over time (the one exception is item QS7 with factor loadings of .82 and .68 at time point 1 and time point 2, respectively).

Both models indicate an adequate fit to the data (Bentler & Bonett, 1980; Browne & Cudeck, 1993; Byrne, 1994; Hu & Bentler, 1995; MacCallum, Browne, & Sugawara, 1996) and provide support for the theoretical model (Table 4). This demonstrates the factorial structure of the SET37_QS, which is one aspect of construct validity (Furr & Bacharach, 2013). The $\chi^2$-tests of exact fit are statistically significant, whereas the objective is to achieve non-significant $p$ values. However, several authors have indicated (e.g., Hatcher, 1994) that a statistically significant $\chi^2$ does not make a confirmatory factor analysis model inadequate, especially when the sample size exceeds 200. Besides, the normed $\chi^2$-values lie within acceptable limits as its value is not higher than 5 (Schumacker & Lomax, 2004).

**Table 4**
Model fit for CFA models from time point 1 (t1) and time point 2 (t2).

| | $\chi^2$ | df | CFI | RMSEA | SRMR |
|---|---|---|---|---|---|
| Models | | | | | |
| t1 | 161.17 | 27 | .96 | .07 | .04 |
| t2 | 127.75 | 27 | .97 | .06 | .04 |

*Note.* All $\chi^2$-values are statistically significant at $p < .0001$. df: degrees of freedom; CFI: comparative fit index; RMSEA: root mean square error of approximation; SRMR: standardized root mean square residual.

**Table 5**
Results from a G Study on the reliability of the SET37_QS instrument for SET (641 students, 9-item instrument, two occasions).

| | df | SS | MS | Variance | Proportion |
|---|---|---|---|---|---|
| ANOVA table | | | | | |
| P | 640 | 4633.04 | 7.24 | .33 | .35 |
| I | 8 | 147.42 | 18.43 | .01 | .01 |
| T | 1 | 7.29 | 7.29 | .00 | .00 |
| P × I | 5120 | 3865.81 | .76 | .20 | .21 |
| P × T | 640 | 529.27 | .83 | .05 | .06 |
| I × T | 8 | 11.02 | 1.38 | .00 | .00 |
| P × I × T | 5120 | 1798.43 | .35 | .35 | .37 |
| Error variances | | | | | |
| Relative | .07 | | | | |
| Absolute | .07 | | | | |
| G-coefficients | | | | | |
| G | .83 | | | | |
| Phi | .83 | | | | |

*Note.* df: degrees of freedom; SS: Sum of Squares; MS: Mean Square; P: persons variance component; I: variance component for items; T: variance component for occasions.

*Reliability analysis (G study)*

To estimate the reliability of the SET37_QS, we use a two-facet fully crossed design (Items (I) × Time (T) × Persons (P)) which includes the obtained measurements from one instrument (I, i.e., the 9-item SET37_QS) on two different occasions (T, i.e., time points 1 and 2) completed by one group of respondents (P, i.e., the 641 students who evaluated a course twice). In the analysis, I & T were treated randomly. Since the outcomes of the SET measurements are commonly used to compare instructional quality (between teachers, or on intra-individual basis between previous and current performance) without using (absolute) cut-off scores, a relative decision was desired. For the analysis, the G1.sps SPSS program for generalizability theory analysis provided by Mushquash and O'Connor (2006) was run.[4] The results are set out in Table 5.

The persons variance component, which is the largest estimated component ($P = .33$), is the variation in the students' answers (i.e., the mean score) over all items and occasions. This variability is desirable (and should not be considered measurement error) since it reflects differences between individual students in their perceptions of the quality of a course. The variance components for items (I), occasions (T) and the interactions between items and occasion (I × T) are estimated close to 0, which suggests that there is almost no variation in the degrees to which the different items in the questionnaire measure the quality of a course, and that SET scores remained consistent over the two occasions. The remaining variance components in their turn however are indicators of error variance. The rather small interaction between persons and occasions (P × T) reveals that SET scores from individual students only slightly differ over occasions. The interaction between persons and items, however, explains about 21% variance and clearly stands out compared to the other variance components. It seems that the items work somewhat inconsistently across persons. The larger three-way interaction (P × I × T) could reflect the variance between persons, items and occasions, but can also be seen as the residual, being influenced by other facets that are not included in the design (Mushquash & O'Connor, 2006). Still, the tree-way interaction (P × I × T) and random error are confounded and, thus, cannot be disentangled (Raykov & Marcoulides, 2011). The error variances presented in Table 5 are indicators of all error variances in the design (absolute), and the error variances for all components that involve both respondents and one other facet

[4] This program can be downloaded from http://flash.lakeheadu.ca/~boconno2/gtheory/gtheory.html.

Please cite this article in press as: P. Spooren, et al.. Assessing the validity and reliability of a quick scan for student's evaluation of teaching. Results from confirmatory factor analysis and G Theory. *Studies in Educational Evaluation* (2014), http://dx.doi.org/10.1016/j.stueduc.2014.03.001

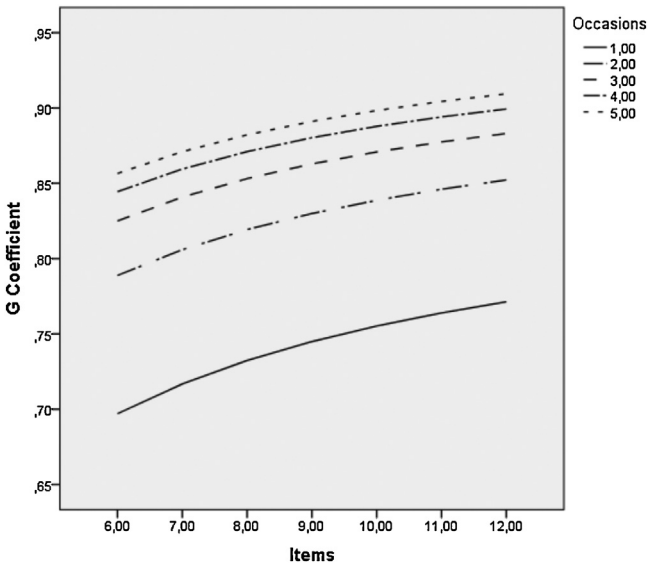


**Fig. 1.** SPSS plot of D Study G Coefficients for the SET37_QS.

(relative). Both measures reflect little error variance in the model. The relative and absolute *G coefficients* indicate that SET scores are reliable across all items and occasions as they exceed the conventional criteria for reliability (usually ranging between .70 and .80).

Still, these coefficients count for SET scores based on two occasions, while SET are usually administered at only one occasion (for instance, at the end of the semester). It is therefore important to estimate the G coefficients by means of a *D Study*, which reveals the G coefficients (and error variances) for different numbers of items and occasions. Fig. 1 shows the different possibilities based on the results of the present G Study (9 items, 2 occasions).

The D Study results indicate that the reliability of the SET37_QS is acceptable when used on only one occasion as well, as this would involve a G coefficient of .75. Decreasing the number of items to six at the same however would cause a drop under the .70 limit. The figure also shows that increasing the number of occasions and the number of items would lead to better G coefficients, although the improvements are much less beyond three occasions and with each item added.

## Discussion and conclusion

The above presented study supports for the use of the SET37_QS questionnaire as a valuable diagnostic instrument for gathering student feedback in internal practices and procedures aimed at both monitoring and improving the quality of instruction in higher education. Starting from the recent observation that various dimensions of effective teaching as measured in SET instruments are influenced by an underlying global factor (i.e., a general instructional skill), we constructed a quick scan instrument derived from the thoroughly validated SET37 questionnaire for SET. From each of six scales in this instrument one item was selected, next to three additional items.

These items were considered to be direct measures of such a global factor and were put through several validity and reliability tests, including interitem analyses, factor analyses, and a G study. The results show that the instrument is acceptably reliable and that the various items indeed are indicators of one global factor reflecting instructional quality.

Still, the present study has some important limitations. First, we still lack a theoretical framework concerning effective teaching upon which SET-instruments can be built (Onwuegbuzie et al., 2009; Penny, 2003). This, of course, has serious consequences regarding the content validity of SET instruments. Although the SET37_QS went in for various validation procedures, it stems not from a general (i.e., inter-institutional or international) agreement on the concept of effective teaching in higher education. This makes the SET37_QS nothing more or less than a well-designed institutional SET questionnaire. Cross-validation procedures in other institutions are needed to prove the generalizability of the instrument in other settings. Second, the rather small numbers of courses ($N = 16$) and disciplines ($N = 6$) that were used in the validation procedure did not allow us to take into account the hierarchical structure of the data (students are nested in courses in disciplines) in the most appropriate way. Future validation procedures should use larger samples to confirm the factor structure of the instrument in a multilevel design. As we are not aware of any examples of G-studies that take into account nested data, we were not able to include the complex data structure in our analysis. The reported G coefficients are, thus, not corrected for the course level. Third, the G study shows that the interaction between persons and items explains about 21% variance, suggesting that the items work somewhat inconsistently across persons. This variance component cannot be neglected as it reflects different relative standings of persons across items (Shavelson & Webb, 2006) and should be re-evaluated in future validation procedures.

Since SET are commonly used in universities and university colleges for both formative and summative purposes, thorough validation procedures for the instruments used are no luxury, even when policy-makers are only interested in concise student feedback. Although we are aware that the SET37_QS will not be applicable to all types of courses nor to all other institutions (due to, for instance, different conceptions on teaching and learning), we hope that this study may be helpful in stimulating other educational researchers and SET practitioners to take at least some pains over investigating the psychometric properties of their instruments by, perhaps, using the several steps that were presented in this study.

## References

Apodaca, P., & Grad, H. (2005). The dimensionality of student ratings of teaching: Integration of uni- and multidimensional models. *Studies in Higher Education, 30*, 723–748 http://dx.doi.org/10.1080/03075070500340101.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606 http://dx.doi.org/10.1037/0033-2909.88.3.588.

Blackmore, J. (2009). Academic pedagogies, quality logics and performative universities: Evaluating teaching and what students want. *Studies in Higher Education, 34*, 857–872 http://dx.doi.org/10.1080/03075070902898664.

Brennan, P. (2001). *Generalizability theory*. New York: Springer.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Burdsal, C. A., & Harrison, P. D. (2008). Further evidence supporting the validity of both a multidimensional profile and an overall evaluation of teaching effectiveness. *Assessment & Evaluation in Higher Education, 33*, 567–576 http://dx.doi.org/10.1080/02602930701699049.

Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows*. Thousand Oaks, CA: Sage Publications.

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education, 44*, 495–518 http://dx.doi.org/10.1023/A:1025492407752.

Chen, Y., & Hoshower, L. (2003). Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment and Evaluation in Higher Education, 28*, 71–88 http://dx.doi.org/10.1080/02602930301683.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Douglas, J., & Douglas, A. (2006). Evaluating teaching quality. *Quality in Higher Education, 12*, 3–13 http://dx.doi.org/10.1080/13538320600685024.

Furr, R. M., & Bacharach, V. R. (2013). *Psychometrics: An introduction*. Los Angeles, CA: Sage.

Ginns, P., & Barrie, S. (2004). Reliability of single-item ratings of quality in higher education: A replication. *Psychology Reports, 95*, 1023–1030 http://dx.doi.org/10.2466/pr0.95.3.1023-1030.

Ginns, P., & Barrie, S. (2009). Developing and testing a student-focussed teaching evaluation survey for university instructors. *Psychological Reports, 104*, 1019–1032 http://dx.doi.org/10.2466/pr0.104.3.1019-1032.

Please cite this article in press as: P. Spooren, et al.. Assessing the validity and reliability of a quick scan for student's evaluation of teaching. Results from confirmatory factor analysis and G Theory. *Studies in Educational Evaluation* (2014), http://dx.doi.org/10.1016/j.stueduc.2014.03.001

Ginns, P., Prosser, M., & Barrie, S. (2007). Students' perceptions of teaching quality in higher education: The perspective of currently enrolled students. *Studies in Higher Education, 32*, 603–615 http://dx.doi.org/10.1080/03075070701573773.

Hatcher, L. (1994). *A step-by-step approach to using the SAS system for factor analysis and structural equation modelling*. North Carolina: SAS Institute.

Hox, J. J. (2010). *Multilevel analysis. Techniques and applications* (2nd ed.). Mahwah, NJ: LEA.

Hu, L. T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). Thousand Oaks, CA: Sage.

Jackson, D. L., Teal, C. R., Raines, S. J., Nansel, T. R., Force, R. C., & Burdsal, C. A. (1999). The dimensions of student's perceptions of teaching effectiveness. *Educational and Psychological Measurement, 59*, 580–596 http://dx.doi.org/10.1177/00131649921970035.

Keeley, J., Smith, D., & Buskist, W. (2006). The Teacher Behaviors Checklist: Factor analysis of its utility for evaluating teaching. *Teaching of Psychology, 33*, 84–90 http://dx.doi.org/10.1207/s15328023top3302_1.

Keeley, J., Furr, R. M., & Buskist, W. (2010). Differentiating Psychology students' perceptions of teachers using the Teacher Behavior Checklist. *Teaching of Psychology, 37*, 16–20 http://dx.doi.org/10.1080/00986280903426282.

Kember, D., & Leung, D. (2008). Establishing the validity and reliability of course evaluation questionnaires. *Assessment & Evaluation in Higher Education, 33*, 341–353 http://dx.doi.org/10.1080/02602930701563070.

Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage Publications.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130–149.

Marsh, H. W. (1982). SEEQ: A reliable, valid and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology, 52*, 77–95 http://dx.doi.org/10.1111/j. 2044-8279.1982.tb02505.x.

Marsh, H. W. (1987). Student's evaluations of university teaching: Research findings, methodological issues, and directions for further research. *International Journal of Educational Research, 11*, 253–388 http://dx.doi.org/10.1016/0883-0355(87)90001-2.

Marsh, H. W. (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). New York: Springer.

Marsh, H. W., Muthèn, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling, 16*, 176–439 http://dx.doi.org/10.1080/10705510903008220.

Mortelmans, D., & Spooren, P. (2009). A revalidation of the SET37-questionnaire for student evaluations of teaching. *Educational Studies, 35*, 547–552 http://dx.doi.org/10.1080/03055690902880299.

Muthén, B. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*, 376–398 http://dx.doi.org/10.1177/0049124194022003006.

Muthén, L. K., & Muthén, B. (1998–2010). *MPlus user's guide* (6th ed.). Los Angeles, CA: Muthén & Muthén.

Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior Research Methods, 38*, 542–547 http://dx.doi.org/10.3758/BF03192810.

Nasser, F., & Hagtvet, K. A. (2006). Multilevel analysis of the effects of student and instructor/courses characteristics. *Research in Higher Education, 47*, 559–590 http://dx.doi.org/10.1007/s11162-005-9007-y.

Onwuegbuzie, A. J., Daniel, L. G., & Collins, K. M. T. (2009). A meta-validation model for assessing the score-validity of student teaching evaluations. *Quality & Quantity, 43*, 197–209 http://dx.doi.org/10.1007/s11135-007-9112-4.

Onwuegbuzie, A. J., Witcher, A. E., Collins, K. M. T., Filer, J. D., Wiedmaier, C. D., & Moore, C. W. (2007). Students' perceptions of characteristics of effective college teachers: A validity study of a teaching evaluation form using a mixed-methods analysis. *American Educational Research Journal, 44*, 113–160 http://dx.doi.org/10.3102/0002831206298169.

Ory, J. C. (2001). Faculty thoughts and concerns about student ratings. *New Directions for Teaching and Learning, 87*, 3–15 http://dx.doi.org/10.1002/tl.23.

Penny, A. R. (2003). Changing the agenda for research into student's views about university teaching: Four shortcomings of SRT research. *Teaching in Higher Education, 8*, 399–411 http://dx.doi.org/10.1080/13562510309396.

Penny, A. R., & Coe, R. (2004). Effectiveness of consultation on student ratings feedback: A meta-analysis. *Review of Educational Research, 74*, 215–253 http://dx.doi.org/10.3102/00346543074002215.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York: Routledge.

Ramsden, P. (1991). A performance indicator of teaching quality in higher education: The Course Experience Questionnaire. *Studies in Higher Education, 16*, 129–150 http://dx.doi.org/10.1080/03075079112331382944.

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment and Evaluation in Higher Education, 30*, 387–415 http://dx.doi.org/10.1080/02602930500099193.

Shavelson, R. J., & Webb, N. M. (2006). Generalizability theory. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of Complementary Methods in Education Research* (pp. 309–322). Mahwah, NJ: Lawrence Erlbaum Associates.

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates.

Spooren, P. (2010). On the credibility of the judge. A cross-classified multilevel analysis on student evaluations of teaching. *Studies in Educational Evaluation, 36*, 121–131 http://dx.doi.org/10.1016/j.stueduc.2011.02.001.

Spooren, P. (2012). *The unbearable lightness of student evaluations of teaching in higher education. A series of studies on their use and validity*. (Unpublished doctoral dissertation) University of Antwerp: Belgium.

Spooren, P., Mortelmans, D., & Denekens, J. (2007). Student evaluation of teaching quality in higher education. Development of an instrument based on 10 Likert scales. *Assessment and Evaluation in Higher Education, 32*, 667–679 http://dx.doi.org/10.1080/02602930601117191.

Spooren, P., Mortelmans, D., & Van Loon, F. (2012). Exploratory Structural Equation Modelling (ESEM): Application to the SET-37 questionnaire for students' evaluation of teaching. *Procedia of Social and Behavioral Sciences, 69*, 1282–1288 http://dx.doi.org/10.1016/j.sbspro.2012.12.063.

Spooren, P., & Van Loon, F. (2012). Who participates (not)? A non-response analysis on students' evaluations of teaching. *Procedia of Social and Behavioral Sciences, 69*, 990–996 http://dx.doi.org/10.1016/j.sbspro.2012.12.025.

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction, 28*, 1–11 http://dx.doi.org/10.1016/j.learninstruc.2013.03.003.

Wanous, J. P., & Hudy, H. J. (2001). Single-Item reliability: A replication and extension. *Organizational Research Methods, 4*, 361–375 http://dx.doi.org/10.1177/109442810144003.

Webb, N. M., Rowley, G. L., & Shavelson, R. J. (1988). Using generalizability theory in counseling and development. *Measurement and Evaluation in Counseling & Development, 21*, 81–90.

Wilson, K. L., Lizzio, A., & Ramsden, P. (1997). The development, validation and application of the Course Experience Questionnaire. *Studies In Higher Education, 22*, 33–53 http://dx.doi.org/10.1080/03075079712331381121.

Zabaleta, F. (2007). The use and misuse of student evaluation of teaching. *Teaching in Higher Education, 12*, 55–76 http://dx.doi.org/10.1080/13562510601102131.

**Pieter Spooren** holds master's degrees in Educational Sciences and Quantitative Analysis in the Social Sciences and a PhD in Social Sciences. He is affiliated as an educational advisor at the Faculty of Political and Social Sciences from the University of Antwerp (Belgium). His particular activities are educational innovation and evaluation of the educational process and of educators. His main research interests focus on students' evaluation of teaching (SET), in particular their use and validity.

**Dimitri Mortelmans** is a professor at the University of Antwerp. He is head of the Research Center for Longitudinal and Life Course Studies (CELLO). He publishes in the domain of family sociology and sociology of labor. Important topics of his expertise are aging, divorce and gender differences in career trajectories.

**Wim Christiaens** holds a master's degree in 'Educational Sciences' and works as an Educational Advisor and a Researcher at the Faculty of Social and Political Sciences from the University of Antwerp (Belgium). His research topics include the evaluation of teaching by students in higher education, and the success rate of first year university students.

Please cite this article in press as: P. Spooren, et al.. Assessing the validity and reliability of a quick scan for student's evaluation of teaching. Results from confirmatory factor analysis and G Theory. *Studies in Educational Evaluation* (2014), http://dx.doi.org/10.1016/j.stueduc.2014.03.001