Establishing the validity and reliability of course evaluation questionnaires

Article *in* Assessment & Evaluation in Higher Education · August 2008

DOI: 10.1080/02602930701563070

Doris Y P Leung
The Hong Kong Polytechnic University
131 PUBLICATIONS **3,572** CITATIONS

Assessment & Evaluation in Higher Education
Vol. 33, No. 4, August 2008, 341–353

Establishing the validity and reliability of course evaluation questionnaires

David Kember[a]* and Doris Y.P. Leung[b]

[a]*The Chinese University of Hong Kong;* [b]*The University of Hong Kong*

This article uses the case of designing a new course questionnaire to discuss the issues of validity, reliability and diagnostic power in good questionnaire design. Validity is often not well addressed in course questionnaire design as there are no straightforward tests that can be applied to an individual instrument. The authors propose the technique of establishing validity by deriving constructs from naturalistic qualitative research—in this case by interviewing award-winning teachers about their principles and practices. Analysis of the interview transcripts led to nine principles of good teaching, which were developed into nine questionnaire scales. Reliability was tested with Cronbach's alpha and with confirmatory factor analysis, as the use of Cronbach's alpha alone can mask issues of multi-dimensionality in scales. The concept of diagnostic power as the ability of an instrument to distinguish between related constructs is introduced. This is important in course evaluation questionnaires, as it enables relative strengths and weaknesses to be identified, which makes it possible to advise on remedial action.

Questionnaire design

A substantial part of social science research makes use of questionnaires to gather data. Universities, like many other types of organizations, use them widely to seek opinions from stakeholders on a wide range of issues. Many of these questionnaires are well designed; so gather useful and accurate information for the intended purpose. Others are not.

This article considers the issue of good questionnaire design. The focus is on reliability and validity, since these are the two criteria most widely used to determine whether or not an instrument is usable. The ability of a questionnaire to fulfil its intended purpose of providing useful feedback and information relating to its intended application is also considered.

The discussion concentrates on course evaluation questionnaires, which are the most widely used instruments in higher education. It is important that they are well designed, since they are used in quality assurance procedures to ensure that teaching and learning are of a reputable standard. They also find an increasing use in personnel decisions, as an integral component of staff appraisal procedures. Design concerns are raised through consideration of the development and testing of a new course evaluation questionnaire at one university. The discussion relating to validation, reliability and diagnostic power should be applicable to other forms of questionnaire.

*Corresponding author. Email: david.kember@cuhk.edu.hk

ISSN 0260-2938 print/ISSN 1469-297X online

DOI: 10.1080/02602930701563070
http://www.informaworld.com

Validity

Of the numerous course or teaching evaluation questionnaires, few appear to have satisfactory evidence of validity. There are readily available statistics to measure reliability but no simply computed statistic that establishes validity. Validity is established if an instrument actually provides a measure of what it purports to measure. In the case of a course evaluation questionnaire, the construct in question is the quality of learning and teaching in a course.

To provide an appropriate measure a course questionnaire must be based on a valid model of quality in education. This then poses a fundamental issue, because there has been considerable argument over a definition of quality in education. Just about the only agreement is that the topic is complex and that many formulations of quality are possible. Marsh (1987) starts the chapter on validity, in his definitive review of research into students' evaluations of university teaching, by noting that 'Student ratings, which constitute one measure of teaching effectiveness, are difficult to validate since there is no single criterion of effective teaching' (285).

Approaches to establishing validity

While the issue of validity is often ducked, there are several approaches that are used. Non-specialists often rely on face validity, in which the wording of items in a scale makes some reference to what is being measured. Content validity goes somewhat further by seeking to include the range of facets of a construct and to do so in a balanced way (Moser and Kalton 1979). The obvious problem here is the criteria for accepting validity claims, particularly with a construct as hard to define as good teaching. One approach has been the use of expert panels. However, these are most likely to be composed of those with similar paradigmatic beliefs to the instrument designer, so radical challenges to a questionnaire are unlikely.

Marsh (1987) argues that construct validity is the best practical method available for validating course evaluation instruments. The most common design compares students' evaluation ratings with other measures of teaching such as instructor-self evaluation, peer rating, rating by external observers and ratings with other instruments.

In spite of extensive testing of construct validity, many of the course evaluation instruments most often cited in the literature have been criticized as being consistent with teacher-centred models of teaching (e.g. Centra 1993; d'Apollonia and Abrami 1997; McKeachie 1997), and quite inconsistent with learner-centred forms (Kolitch and Dean 1999). This arises because they were often based on early work on instructor evaluation, which had largely positivist origins. Marsh (1987) attributes the origins of the dimensions in typical questionnaires to 'a logical analysis of the content of effective teaching and the purposes of students' evaluations, supplemented perhaps with literature reviews' (264). A set of 19 dimensions by Feldman (1976) is cited as being influential and an example of best practice.

Examination of Feldman's dimensions clearly shows the nature of the model used to derive the dimensions. No less than 11 of the 19 dimension labels commence with the word 'Teacher's'. Of the remainder, two are concerned with the clarity of presentations and two with materials. The model is of the teacher-centred content-oriented type. The dimensions fit well with didactic teaching, but it is hard to see the applicability of many of the dimensions to other more student-centred forms of teaching.

The difficulties with construct validity, with regard to course evaluation questionnaires, lie in cross-validation of instruments or measures based on similar theoretical origins. Course evaluation questionnaires have most frequently been validated against other measures emanating from behavioural research with teacher-centred models of good teaching. Alternatively they utilize different types of raters, such as students and instructors, using the same instrument. As they are based on related theoretical origins correlations between measures are inevitably high, so claims of validity are made. The instruments and what they are compared with all measure the same teacher-centred model of good teaching. However, if the teacher-centred model of good teaching is rejected, and there are compelling grounds for doing so, the procedures establish that the measures compared are equally valid measures of poor teaching. Alternatively, neither is a valid measure of good teaching.

There are several grounds for not accepting teacher-centred models as appropriate descriptions of good teaching. One compelling line of evidence is the research which shows that teachers with teacher-centred and content-oriented beliefs tend to induce surface approaches to learning. Gow and Kember (1993) and Kember and Gow (1994) established this at the departmental level, while Trigwell et al. (1999) drew the same conclusion in a study of individual teachers.

Establishing validity by deriving constructs from naturalistic research

In this article we argue for a form of construct validation which avoids this problem by eschewing hypothesis-based or top-down generation of the theory underpinning the design of an instrument. Essentially it moves the origins of course questionnaire design from a positivist paradigm to a naturalistic base. Rather than designing a questionnaire from the theory of the researcher, the constructs are derived in an open way from the perspectives of the subjects. In our case the model of good teaching is grounded in the beliefs and practices of the teachers judged to be the best in the university.

This approach to establishing validity is consistent with Messick's (1996) argument that directness and authenticity are important constructs which should be taken into account when validity is considered. Messick (1992) further asserted that validity is an integrated evaluative judgement of the extent to which evidence supports the appropriateness of both interpretations and actions.

In developing a course questionnaire at a university in Hong Kong this issue of validity was tackled by basing the design on research into the practices of teachers in the university who were judged to be exemplary. Eighteen teachers awarded the Vice-Chancellor's Award for Exemplary Teaching, in the first three years of its operation were interviewed on the topic of their beliefs and practices as a teacher. Analysis of the transcripts using grounded theory (Glaser and Strauss 1967; Lincoln and Guber 1985) and the constant comparative method (Strauss and Corbin 1990), found a set of common principles of what constitutes excellence in teaching (Kember with McNaught 2007). While there was variety in the way these were applied, there was a very high degree of consistency in holding to the principles.

The analysis of the qualitative data was confirmed by a rigorous set of verification procedures. These were designed to ensure that the practices and beliefs of the exemplary teachers were genuinely enshrined in the set of principles and the accompanying exposition. Semi-quantitative searches of the transcripts were used to show the proportion of the interviewees mentioning important constructs. These were generally very high but not seen as exhaustive, because interviews can omit a topic, even if it is consistent with the views of the interviewee. There were two levels of verification with the 18 teachers. First, each was asked to verify the transcript of his/her interview. Second, every interviewee was asked to comment on each part of the completed analysis to verify that the principles and explanation were consistent with his/her views and practices.

Table 1 shows the set of principles of excellent teaching (adapted from Kember with McNaught 2007). All principles were used in designing the course questionnaire except for

Table 1. The principles of excellent teaching used as a framework.

1. Teaching and curriculum design needs to be consistent with meeting students' future needs. This implies the development of a range of generic capabilities including:
    - self-managed learning ability
    - critical thinking
    - analytical skills
    - team-work
    - leadership
    - communication skills
2. Ensure that students have a thorough understanding of fundamental concepts, if necessary at the expense of covering excessive content.
3. Establish the relevance of what is taught by:
    - using real-life examples
    - drawing cases from current issues
    - giving local examples
    - relating theory to practice.
4. Challenging beliefs is important to:
    - establish appropriate ways of learning and beliefs about knowledge
    - deal with misconceptions of fundamental concepts
5. Meaningful learning is most likely to occur when students are actively engaged with a variety of learning tasks. Discussion is an important learning activity
6. Establishing empathetic relationships with students is a prerequisite to successful interaction with them. To do this you need to know them as individuals
7. Good teachers accept that it is their responsibility to motivate students to achieve the high expectations they have of them. Motivation comes through:
    - encouraging students
    - the enthusiasm of the teacher
    - interesting and enjoyable classes
    - relevant material
    - a variety of active learning approaches
8. Planning programmes and courses involves consideration of students' future needs. The plans ensure that aims, fundamental concepts, learning activities and assessment are consistent with achieving outcomes related to future student needs. Feedback needs to be gathered to inform each of these elements in the curriculum design process
9. Thorough planning is needed for each lesson, but plans need to be adapted flexibly in the light of feedback obtained in class
10. Assessment must be consistent with the desired learning outcomes and eventual student needs if these are to be achieved. Assessment should, therefore, comprise authentic tasks for the discipline or profession

Source: Adapted from: Kember, D. with McNaught, C. 2007. *Enhancing university teaching: lessons from research into award winning teachers* (Abingdon, Routledge).

the first. This is about the development of generic capabilities, which is normally considered to be a programme-level goal. It is, therefore, more appropriate to evaluate the development of generic capabilities at the university or programme level, rather than during course evaluation.

The remaining nine principles were then used to derive a set of dimensions of good teaching. In a questionnaire each dimension becomes a scale, made up of a number of items that deal with important facets of the dimension. In this case the principles serve to identify scales and provide guidance in wording the items. By following this path the questionnaire becomes grounded in the constructs derived from the interviews. If these are accepted as forming a valid model of good teaching in the university, then it follows that the questionnaire has the same construct validity. Authenticity and applicability are also dealt with since the base data were gathered from exemplary teachers from all faculties within the university for which the questionnaire was designed.

An initial trial version of the Exemplary Teacher Course Questionnaire (ETCQ) with 49 items was developed. All the scales contained between five and seven items. This was more than were likely to feature in the final version. The testing process could then be used to test which items had better psychometric properties in respect of a particular scale (Noar 2003), so that less satisfactory items could be removed. Given the concern about the reluctance of students to complete questionnaires, the aim has to be to produce an instrument that is as short as possible, while at the same time including all relevant dimensions and providing a reliable measure of them.

The scales and items for the initial test version of the questionnaire are listed in Table 3 (see below). Comparison between Tables 1 and 3 shows clearly the grounding of the dimensions and items in the principles of excellent teaching. Wording of items was also informed by the detailed analysis of the interviews.

Reliability

Having decided on a set of dimensions, it is then necessary to ensure that they form a reliable scale when administered to a particular population. The most commonly used measure of reliability is Cronbach's alpha coefficient (e.g. Raykov and Shrout 2002). This gives a measure of the accuracy or consistency with which a set of items measures a single construct (Miller 1995). When instruments are described in journal articles, figures for Cronbach's alpha are normally given. Computing these is straightforward. SPSS for example has a procedure called 'reliability' (Norusis 2002).

It might be thought that there would be little left to discuss regarding a statistic which was invented a long time ago (Cronbach 1951) and very widely used. However, papers are still being written about Cronbach's alpha. There are two issues that concern the design of course evaluation questionnaires which will be raised here.

The first is that Cronbach's alpha values are a function of the number of items in a scale (see for example, Nunnally 1978; Miller 1995). Adding more items to a scale tends to increase alpha. The problem is that students can be asked to complete so many questionnaires nowadays that many have become reluctant to complete lengthy ones. Questionnaire design, therefore, has to compromise between reliability and length (Scriven 1994).

Cronbach's alpha is also a function of the average inter-correlation between items (Green et al. 1977). This means that higher alpha values are obtained if it is assumed that the construct being measured is unidimensional. However, education—and the social sciences generally—deal with constructs which are complex. There can then be a tension between fully describing a construct and achieving reliable measurements. Including all pertinent facets of a construct in a scale will result in multidimensionality, which will reduce alpha values. Restricting the number of dimensions in a scale will increase alpha values, but will mean that the scale no longer addresses the complexity of the construct. The dichotomy is between measurement and validity.

The most common practice with verification of questionnaires, of restricting analysis and reporting to Cronbach's alpha values, masks this issue because the statistic does not reveal the dimensionality of the scale (McDonald 1981; Hattie 1985). We advocate the use of structural equation modelling (SEM) to address this issue as it can reveal the underlying dimensionality of a scale (Rubio et al. 2001; Noar 2003). Those unfamiliar with SEM would find that Byrne (1994) provides a good introduction. In the development of the questionnaire discussed in this article the reliability tests and structural equation modelling did not lead to differing conclusions. However, in the development of a revised version of the Learning Process Questionnaire (Kember et al. 2004), the use of structural equation modelling proved to be vital because of the hierarchical multidimensional character of the learning approach scales.

Testing

Data for the pilot test came from 662 Hong Kong students studying full-time associate degrees in 20 courses in a variety of disciplines. The students were asked to complete the ETCQ questionnaire during their class. The responses to each of the 49 items were gathered with a five-point Likert scale ranging from 'strongly agree' to 'strongly disagree'. The sampling could be interpreted as a convenience sample, but any minor differences from randomness would matter little since the study was aiming to test an instrument, rather than examine the attributes of a population.

Analysis and results

Reliability for the full version

In line with the discussion above, the scales of the full version of the questionnaire were assessed for reliability by computing Cronbach's alpha values with the 'reliability' procedure of SPSS11.0 (Norusis 2002). The values of alpha for the nine scales ranged from 0.79 for Understanding Fundamental Concepts to 0.91 for Flexibility, as shown in the second column of Table 2. All values comfortably exceeded normally accepted values for a scale to be considered reliable (Schmitt 1996).

Confirmatory factor analysis of the full version

Confirmatory factor analysis (CFA) was used to test the structure of each scale and the dimensionality of the instrument. The hypothesized model therefore had nine factors corresponding to the nine scales. It was hypothesized that each factor correlated with each other. The items for each scale function as indicators for the respective factor.

Table 2. Cronbach's alpha values for the full and the final reduced versions of the nine scales of the Exemplary Teacher Course Questionnaire.

	Full version		Three-item version	
Scale	Alpha values	No. of items	Alpha values	No. of items
Understanding fundamental content	0.79	5	0.76	3
Relevance	0.83	5	0.82	3
Challenging beliefs	0.84	6	0.77	3
Active learning	0.87	5	0.87	3
Teacher–student relationships	0.91	5	0.88	3
Motivation	0.89	6	0.87	3
Organization	0.91	7	0.89	3
Flexibility	0.91	5	0.87	3
Assessment	0.86	5	0.79	3

The fit of the model to the data was tested with the EQS package (Bentler 1995). Assessment of model fit was based on a goodness-of-fit index, namely the Comparative Fit Index (CFI; Bentler 1990). Two absolute misfit indices were used: the Root Mean Square Error of Approximation (RMSEA; Browne and Cudeck 1993) and the standardized root mean squared residual (SRMR; Bentler 1995). A model with SRMR < 0.08, RMSEA < 0.06 and CFI > 0.95 would be considered as an excellent fit to the data.

The hypothesized model provided a barely adequate fit to the observed data as suggested by the goodness-of-fit-indices (SRMR = 0.048, RMSEA = 0.057, CFI = 0.896). All the factor loadings were statistically significant, ranging from 0.56 to 0.91. Correlations among the nine factors varied from 0.67 to 0.92, which is consistent with the theory that the factors are expected to have large positive correlations.

It could be argued that the best questionnaire to use is the full version. Including all items in the scales gives the fullest characterization of the construct, so should provide the most valid representation of the construct. The alpha values were high so reliability can be considered highly acceptable. The fit of the overall multidimensional model of good university teaching to the data was sufficient to provide support for the instrument.

Scale reduction

In spite of these claims in respect of the full instrument, steps were taken to reduce the number of items in each scale because of the practical concern that completion rates tend to drop for lengthier questionnaires. The reduction of items was based on two sets of evidence: use of reliability statistics and a series of confirmatory factor analyses using EQS. Values for Cronbach's alpha for scales were produced with the 'reliability' procedure of SPSS11.0. The statistic gives a value for alpha if an item were deleted, together with inter-item correlations. These statistics give useful guidance in deciding which items to delete.

Confirmatory factor analysis was performed to assess the dimensionality of each of the nine scales of ETCQ. The model tested was that of the scale as a latent factor with each item as an indicator. The results of the test show the degree of fit of the model to the data and show the loading of each item on the factor. Those items with higher loadings make a greater contribution to the scale and are, therefore, generally the better ones to retain. This type of testing can also test alternative relationships between items, sub-scales and scales. This is particularly helpful where multidimensionality is suspected, as the tests of reliability give little indication of this.

Both Gerbing and Anderson (1988) and Bentler (1995) suggested that the standardized residual matrix be examined, as large standardized residuals associated with specific items indicate that the model probably does not explain the covariances among the items adequately. Therefore, those items with large standardized residuals, generally, should be dropped. In this case, this procedure led to the same outcomes as examination of factor loadings, so the residual matrices are not displayed.

The first column of Table 3 gives the alpha value for the scale if that item were to be removed. A process of successive reduction of a single item followed by re-computation was then performed. However, the outcome was identical to rejecting the least reliable items in the initial computation. Table 3 therefore only shows the initial computation. The factor loadings are also shown in Table 3. These are the loading of the item on the scale as a latent factor in the CFA analyses. It is noticeable that there is complete consistency between the alpha values and the factor loadings. The items making the least contribution to reliability have the lowest factor loadings.

Table 3. Values of Cronbach's alpha if the item is removed, and factor loadings for the nine scales in the original version of the Exemplary Teacher Course Questionnaire.

	α if item removed	Factor loadings
Understanding Fundamental Concepts		
1 This course concentrated on fundamental concepts	**0.76**	**0.66**
2 In each class the key points were made clear	**0.75**	**0.71**
3 We were not overloaded with a lot of facts	0.76	0.61
4 In this course I learnt the key principles	**0.73**	**0.76**
5 The amount of content covered was not excessive	0.77	0.56
Relevance		
6 I could understand the relevance of what was taught	0.81	0.58
7 Theory was related to practical applications	0.81	0.58
8 Local examples were used to show the relevance of material	**0.78**	**0.75**
9 I could see the relevance of material because real-life examples were given	**0.76**	**0.84**
10 Current issues were used to make the course interesting	**0.79**	**0.74**
Challenging Beliefs		
11 In this course we were exposed to different points of view	0.81	0.70
12 After taking this course I have a better understanding of fundamental concepts	**0.81**	**0.71**
13 I have become more flexible in my learning	**0.81**	**0.71**
14 There were times when the teacher(s) made us think deeply about important issues	0.81	0.70
15 I found this course challenging	0.83	0.58
16 I am now more willing to change my views and accept new ideas	**0.81**	**0.70**
Active Learning		
17 A variety of teaching methods were used	0.89	0.63
18 Students were given the chance to participate in class	**0.85**	**0.81**
19 There were activities which encouraged the application of knowledge	0.86	0.77
20 There was discussion between students in class	**0.85**	**0.84**
21 The teaching staff promoted discussion in class	**0.85**	**0.86**
Teacher–Student Relationships		
22 I feel that our teaching staff understood our learning needs	0.90.	0.73
23 There was a friendly relationship between teaching staff and students	**0.88**	**0.9**
24 The communication between teaching staff and students is good	**0.87**	**0.93**
25 Our teacher(s) knew the individuals in the class	**0.89**	**0.74**
26 Our teacher(s) paid attention to the progress of individual students	0.89	0.72
Motivation		
27 The high expectations of this course motivated me to learn	0.88	0.71
28 The teacher(s) were enthusiastic	**0.87**	**0.74**
29 I found the classes enjoyable	**0.86**	**0.87**
30 This was an interesting course	**0.86**	**0.87**
31 The teacher(s) encouraged us to try hard	0.88	0.67
32 This was a demanding course but I learnt a lot from it	0.88	0.70.

Table 3. *(Continued).*

		α if item removed	Factor loadings
Organisation			
33	This course was well organized	**0.89**	**0.86**
34	This course was well planned	**0.89**	**0.87**
35	Each class was well planned	**0.89**	**0.79**
36	The expected learning outcomes of the course were clear	0.89	0.76
37	The objectives of the course were clear	0.89	0.76
38	The learning activities helped us achieve the expected learning outcomes	0.90.	0.67
39	This course catered well to our future needs	0.90.	0.63
Flexibility			
40	I found teaching staff helpful when I had difficulty understanding concepts	**0.89**	**0.82**
41	The teaching staff were sensitive to student feedback	**0.89**	**0.83**
42	The teacher(s) interacted with students in class	0.89	0.82
43	The teacher(s) were helpful when asked questions	**0.88**	**0.85**
44	Teaching was adapted in the light of student feedback	0.90.	0.78
Assessment			
45	The type of assessment related closely to the expected learning outcomes	**0.83**	**0.75**
46	The assessment tested our understanding of key concepts	**0.82**	**0.77**
47	To do well in the course you needed to have good analytical skills	0.83	0.72
48	A variety of assessment methods were used	**0.83**	**0.73**
49	The assessment was a valid test of the course objectives	0.83	0.72

Notes: The items with values of α and factor loadings in bold are those included in the version with three items per scale. The questionnaire is © 2006 David Kember and Doris Y. P. Leung.

The questionnaire was reduced to three items in each scale. The items making the least reliable and least significant contribution to scales were eliminated. The items removed were those with the lowest factor loadings, associated with the largest standardized residuals, and the most detrimental effect on alpha values. The items remaining in the three items per scale version of the questionnaire are indicated in Table 3 in bold in the final two columns.

The Cronbach's alpha values for the final three-item scales are also given in Table 2. The alpha values for each of the nine scales in the reduced three-item version, which ranged from 0.76 to 0.89, were slightly lower than those for the original version. Psychometrically, the reduction of items did not significantly reduce the scales' internal consistency.

Confirmatory factor analysis of the reduced version

The next step in the testing process was to show that the data from the reduced form of the questionnaire showed a good fit to the model envisaged in the design of the instrument. This model consisted of nine constructs, or latent variables, each with three indicators or items. The latent variables were envisaged as dimensions of quality in teaching and learning, and so are discrete entities, though inter-correlated with each other because each is a facet of teaching and learning. Confirmatory factor analysis, using the EQS package, provided a test of the extent to which the data fit to the hypothesized model.

The fit indexes obtained for this model were: CFI = 0.968, SRMR = 0.039 and RMSEA = 0.045. These values indicate that the model is a very good fit to the data, confirming that the reduced version of the questionnaire provides a valid measure of the structure envisaged in its design. In general, the pattern of correlations among the nine factors of the reduced version resembled the original version, varying from 0.53 to 0.91.

Diagnostic power

In carrying out the practical tests on the instrument it became apparent that there was another criterion by which course evaluation questionnaires ought to be assessed, which does not normally appear to be taken into consideration. We have called the criterion *diagnostic power*, which is the degree to which the questionnaire distinguishes between similar constructs in the questionnaire. For a course questionnaire, it is the capability of the instrument to distinguish strengths and weaknesses in a course or teacher. In other words it is the degree of diagnosis and the extent to which results can be used to identify remedial action by pointing out which aspects of teaching need attention.

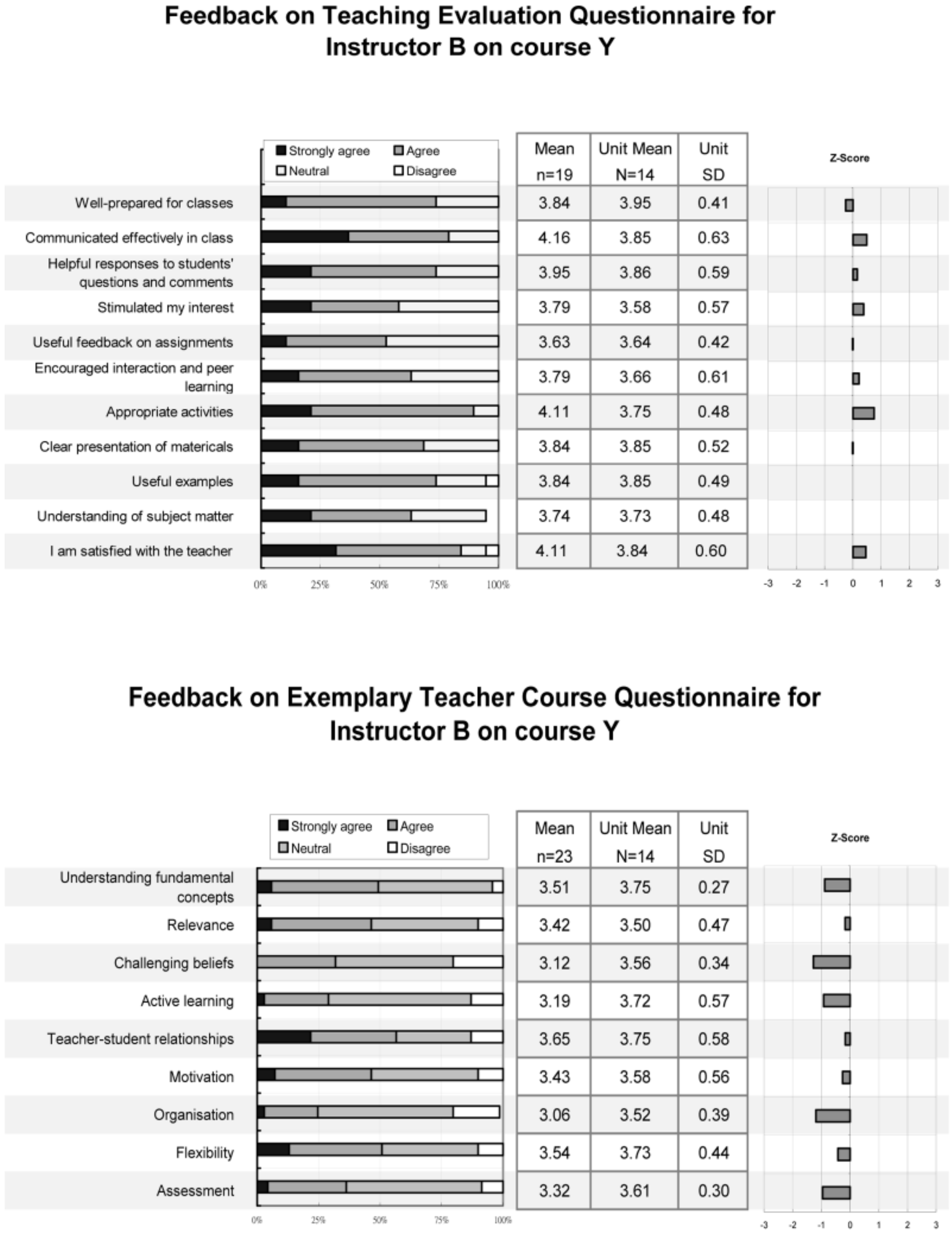
Results from course evaluation questionnaires are commonly reported in comparison with scores from the remainder of a university or faculty. We do this by using z-scores, which are the number of standard deviations from the mean.

The sample for the tests of the instrument had previously completed another questionnaire, the Teaching Evaluation Questionnaire (TEQ), to report on the teaching in the same courses. The TEQ was an internally developed instrument using 11 individual items for constructs, rather than the scales of the instrument described in this article. We had noticed that the z-scores for a course or teacher, from this instrument, were all very close. The instrument could rank teachers according to overall teaching quality but provided little insight into which aspects of teaching might need attention.

This finding was not surprising in view of the factor structure of the instrument. A factor analysis was performed with the 11 items of the TEQ and only one factor was retained, which explained 77.9% of the total variance. The factor loadings obtained ranged from 0.83 to 0.91. This evidence indicates that all 11 items are measuring one single factor, with each item contributing to the factor to a similar extent. The students appear to have responded to the questionnaire by giving an overall judgement on teaching quality, rather than responding to specific facets of teaching. For an instrument to be diagnostic it seems to be important for it to have an explicit set of scales corresponding to the dimensions of teaching.

By comparison, the z-score profiles from the three-item per scale version of our instrument showed greater variation. For particular courses most z-score profiles had predominantly positive or negative scores, but the degree of variation was markedly higher than for the previous instrument. It was therefore possible to diagnose relative strengths and weaknesses; so advice could be given on which aspects of teaching and learning needed attention for an improvement to occur. A comparison of typical z-score profiles from the two instruments is shown in Figure 1.

We propose the use of the degree of variation of the z-scores among scales (or items) as a measure of the diagnostic power of an instrument. Standard deviations would be an appropriate measure here since the mean values of the z-scores among the nine scores for ETCQ and among the 11 items for TEQ for the courses are expected to be comparable. The mean value of the standard deviations across the ETCQ was higher than that for TEQ in 11 of the 14 courses.



	Mean n=19	Unit Mean N=14	Unit SD
Well-prepared for classes	3.84	3.95	0.41
Communicated effectively in class	4.16	3.85	0.63
Helpful responses to students' questions and comments	3.95	3.86	0.59
Stimulated my interest	3.79	3.58	0.57
Useful feedback on assignments	3.63	3.64	0.42
Encouraged interaction and peer learning	3.79	3.66	0.61
Appropriate activities	4.11	3.75	0.48
Clear presentation of matericals	3.84	3.85	0.52
Useful examples	3.84	3.85	0.49
Understanding of subject matter	3.74	3.73	0.48
I am satisfied with the teacher	4.11	3.84	0.60

	Mean n=23	Unit Mean N=14	Unit SD
Understanding fundamental concepts	3.51	3.75	0.27
Relevance	3.42	3.50	0.47
Challenging beliefs	3.12	3.56	0.34
Active learning	3.19	3.72	0.57
Teacher-student relationships	3.65	3.75	0.58
Motivation	3.43	3.58	0.56
Organisation	3.06	3.52	0.39
Flexibility	3.54	3.73	0.44
Assessment	3.32	3.61	0.30

Figure 1. A comparison between z-score profiles from the Teaching Evaluation Questionnaire and the three-item form of the Exemplary Teacher Course Questionnaire from one typical course.

Conclusion

This article has discussed the procedures for establishing the validity and reliability of questionnaires, by using the example of the development of a course evaluation questionnaire. The validity of many course questionnaires in common use has not been established to a satisfactory degree. Face and content validity have obvious limitations. Tests of construct validity are not easy to perform and can be of limited value if the alternative measurements are derived from the same source.

This article advocates the use of naturalistic qualitative research to establish the validity of constructs to be included in a questionnaire. In the case of the ETCQ, interviews were conducted with award-winning university teachers. Analysis of the interviews established a set of principles of good teaching. Scales for each principle were devised in a manner consistent with details from the defining characteristics.

Reliability has proved much easier to establish as there are readily available statistical tests. We advocate the use of confirmatory factor analysis in addition to the most commonly used Cronbach's alpha statistic. The latter can mask the presence of multidimensionality, which may well be present in scales that are authentic representations of complex social science phenomena.

We have also introduced another criterion for judging the utility of a questionnaire, which we have called diagnostic power. This is the ability to provide useful feedback by providing clear measurements which distinguish the components of the overall construct measured by the instrument. In the case of a course evaluation questionnaire it means identifying relative strengths and weaknesses in teaching so that appropriate remedial action can be identified.

Notes on contributors

David Kember is Professor of Learning Enhancement at the Chinese University of Hong Kong. He has previously held academic positions in Hong Kong, Australia, Papua New Guinea, Fiji and the United Kingdom. His research interests are in student learning and action research to improve the quality of teaching.

Doris Y. P. Leung is currently a post-doctoral fellow in the Department of Nursing Studies at the University of Hong Kong. She was previously Research Fellow at the Chinese University of Hong Kong and the Hong Kong Polytechnic University. She specializes in applications of Structural Equation Modeling.

References

Bentler, P.M. 1990. Comparative fit indexes in structural models. *Psychological Bulletin* 107: 238–246.

———. 1995. *EQS: structural equations program.* Encino, CA: Multivariate Software.

Browne, M.W., and R. Cudeck. 1993. Alternative ways of assessing model fit. In *Testing structural equation models,* eds. K.A. Bollen and J.S. Long. Newbury Park, CA: Sage Publications.

Byrne, B.M. 1994). *Structural equation modelling with EQS and EQS/Windows: basic concepts, applications and programming.* Thousand Oaks, CA: Sage Publications.

Centra, J. 1993. *Reflective faculty evaluation.* San Francisco: Jossey-Bass.

Cronbach, L.J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 45: 99–105.

D'Appollonia, S., and P.C. Abrami. 1997. Navigating student ratings of instruction. *American Psychologist* 52: 1198–1208.

Feldman, K.A. 1976. The superior college teacher from the student's view. *Research in Higher Education* 5: 243–288.

Gerbing, D.W., and J.C. Anderson. 1988. An updated paradigm for scale development incorporating unidimensionality and its assessment. *Journal of Marketing Research* 25: 186–192.

Glaser, B.G., and A.L. Strauss. 1967. *The discovery of grounded theory.* Chicago: Aldine.

Gow, L., and D. Kember. 1993. Conceptions of teaching and their relationship to student learning. *British Journal of Educational Psychology* 63: 20–33.

Green, S.B., R.W. Lissitz, and S.A. Mulaik. 1977. Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement* 37: 827–838.

Hattie, J. 1985. Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement* 9, no. 2: 139–164.

Kember, D., J. Biggs, and Leung, D.Y.P. 2004. Examining the multidimensionality of approaches to learning through the development of a revised version of the Learning Process Questionnaire. *British Journal of Educational Psychology* 74, no. 2: 261–280.

Kember, D., and L. Gow. 1994. Orientations to teaching and their effect on the quality of student learning. *Journal of Higher Education* 65, no. 1: 58–74.

Kember, D. with C. McNaught. 2007. *Enhancing university teaching: lessons from research into award winning teachers.* Abingdon: Routledge.

Kolitch, E., and A.V. Dean. 1999. Student ratings of instruction in the USA: hidden assumptions and missing conceptions about ‘good’ teaching. *Studies in Higher Education* 24, no. 1: 27–42.

Lincoln, Y., and E. Guber. 1985. *Naturalistic inquiry.* Newbury Park, CA: Sage Publications.

Marsh, H.W. 1987. Students’ evaluations of university teaching: research findings, methodological issues, and directions for future research. *International Journal of Educational Research* 11: 253–388.

McDonald, R.P. 1981. The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology* 34: 100–117.

McKeachie, W. 1997. Student ratings: the validity of use. *American Psychologist* 52: 1218–1225.

Messick, S. 1992. The interplay of evidence and consequences in the validation of performance assessments. Paper presented at the *Annual Meeting of the National Council on Measurements in Education.* San Francisco, CA: April.

———. 1996. Validity and washback in language testing. *Language Testing* 13, no. 3: 241–256.

Miller, M.B. 1995. Coefficient alpha: a basic introduction from the perspectives of classical test theory and structural equation modelling. *Structural Equation Modeling* 2, no. 3: 255–273.

Moser, C.A., and G. Kalton. 1979. *Survey methods in social investigation.* 2nd edn. Aldershot: Gower.

Noar, S.M. 2003. The role of structural equation modeling in scale development. *Structural Equation Modeling* 10, no. 4: 622–647.

Norusis, M. 2002. *SPSS11.0 Guide to Data.* Upper Saddle River, NJ: Prentice Hall.

Nunnally, J.C. 1978. *Psychometric theory.* 2nd edn. New York: McGraw-Hill.

Raykov, T., and P.E. Shrout. 2002. Reliability of scales with general structure: point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling* 9: 195–212.

Rubio, D.M., M. Berg-Weger, and S.S. Tebb. 2001. Using structural equation modeling to test for multidimensionality. *Structural Equation Modeling* 8, no. 4: 613–626.

Schmitt, M. 1996. Uses and abuses of coefficient alpha. *Psychological Assessment* 8, no. 4: 350–353.

Scriven, M. 1994. Using student ratings in teacher evaluation: evaluation perspective. *Newsletter of the Centre for Research and Educational Accountability and Teacher Evaluation* 4, no. 1: 1–4.

Strauss, A., and J. Corbin. 1990. *Basics of qualitative research: grounded theory procedures and techniques.* Newbury Park, CA: Sage Publications.

Trigwell, K., M. Prosser, and F. Waterhouse. 1999. Relations between teachers’ approaches to teaching and students’ approaches to learning. *Higher Education* 37: 57–70.