

Disciplinary Differences in Student Ratings of Teaching Quality

David Kember · Doris Y. P. Leung

Received: 7 October 2009 / Published online: 27 October 2010
© The Author(s) 2010. This article is published with open access at Springerlink.com

Abstract Comparisons of large samples of course and teaching evaluation questionnaire data show consistent disciplinary differences. The current study examined the disciplinary differences in a theoretical model positing the impact of the perceived nature of teaching and learning environment on the development of generic capabilities by testing of (1) whether a common model of good teaching operated across disciplines and (2) the extent of deployment of teaching variables and their impact on learning outcomes. The sample consisted of 3,305 first and third year Chinese undergraduates of a university in Hong Kong, divided into four broad disciplinary groupings. Multiple-group structural equation modelling analysis showed configural invariance of the hypothesised model, suggesting a common model of good teaching across disciplines; and significant differences in the magnitude of structural paths and latent mean values across the four disciplinary groups were obtained reflecting differences between disciplines in the extent to which elements within the teaching and learning environment were brought into play. Possible reasons in terms of the epistemological nature of the disciplines were given to explain for the disciplinary variations.

Keywords Disciplinary difference · Epistemology · Graduate capabilities · Perceived quality of teaching · Teaching and learning environments

Disciplinary Differences in Teaching

The practice of evaluating teaching and courses in higher education, through Course And Teaching Evaluation Questionnaires (CTEQ) is now widespread to the extent that it is rare

D. Kember
Centre of the Enhancement of Teaching and Learning, The University of Hong Kong,
RR307, Run-Run Shaw Building, Pokfulam, Hong Kong

D. Y. P. Leung (✉)
School of Nursing, The University of Hong Kong, 4/F, William M W Mong Block,
21 Sassoon Road, Pokfulam, Hong Kong
e-mail: dorisl@hkucc.hku.hk

to find colleges which do not routinely evaluate courses. As the practice has become so extensive, a substantial literature has developed. Several reviews have concurred that course evaluation questionnaires give quite reliable and valid measures of teaching effectiveness (Aleamoni 1999; Feldman 1996, 1997, 2007; Marsh 1987, 2007; Marsh and Dunkin 1997; McKeachie 1997).

For all the research, though, there is a well-substantiated finding which does not seem to have been adequately explained. Comparisons of large samples of CTEQ data consistently show disciplinary differences. Teaching in the arts, humanities and social sciences tends to receive higher ratings than engineering and science, with other discipline groupings in intermediate positions. As reviews have concluded that CTEQ data are valid and reliable, this seems to indicate that teaching quality varies between disciplines. There does not seem to be any consensus, however, as to whether this is the case and if so why.

Surveys or reviews of studies on teaching or course evaluation questionnaires commonly show variations in ratings by academic discipline. Feldman (1978) reviewed 11 studies which compared ratings across disciplines and found that the humanities and arts tended to be rated higher than sciences, engineering and business administration. Cashin (1990) analysed large data sets from two widely used course evaluation instruments in the US to find that ratings tended to be higher in the arts and humanities than in science, engineering and business. Barnes and Patterson (1988) also produced closely related findings. There does, therefore, seem to be widespread evidence of disciplinary differences in ratings, with some consistency about which disciplines tend to receive better ratings. Arts, the humanities and the social sciences tended to be in a higher rated group, while science and engineering are on average lower. Other disciplines are intermediate or less consistent.

There is evidence of disciplinary differences on measures of teaching and learning other than CTEQ instruments. For example, both of the most widely used instruments for measuring approaches to learning have indicated disciplinary effects. Entwistle and Ramsden (1983) reported scores for the Approaches to Studying Inventory. English and history were higher on deep approach and lower on surface approach than physics and engineering. Biggs (1987) gathered data from a wide sample of Australian universities with the Study Process Questionnaire. Arts and science students had equal mean scores for deep approach in the first year, but in each subsequent year the arts students had higher mean scores.

The grouping of the disciplines has most commonly been interpreted using Biglan's (1973) classification of disciplines (e.g. Neumann and Neumann 1985). Biglan proposed a $2 \times 2 \times 2$ categorisation scheme for disciplines. The three distinguishing criteria were:

- degree of consensus on paradigm development, hard versus soft;
- presence of practical application, pure versus applied; and
- presence of living organisms, life versus non-life.

The research into disciplinary differences in ratings has most often drawn upon the hard versus soft distinction. This may be because the broad field of disciplinary research has most commonly examined research-related concepts and the behavior of the professoriate (e.g. Becher and Trowler 2001). For undergraduate teaching, particularly in the initial years, these research-related constructs may not be particularly pertinent. For example, applied degrees commonly start by building a solid foundation of pure basic knowledge, leaving application to the final parts (Schön 1987).

The large survey by Cashin (1990) found that the disciplinary differences in ratings explained only a modest percentage of the variance, but of sufficient magnitude that the

effect could not be ignored. As pointed out by Kwan (1999, p. 184) and Marsh (1987, p. 309) the percentage of variance shows the strength of the relationship between variables and the explanatory power. However, a more practical issue is the absolute effect on ratings, because that is what is taken into account when teacher evaluation ratings are used in appraisal judgements. That is better measured by an effect size.

Kwan's (1999) study gives both the percentage of variance and effect sizes on the effects of factors on evaluation ratings. Discipline explained about 10% of the variance. Effect sizes for greatest differences between five discipline groupings were computed for each of the six scales in the particular CTEQ instrument. They ranged from 0.43 to 0.86 which means that disciplinary differences in CTEQ ratings have to be treated as quite appreciable effects according to the recommendation by Cohen (1988).

Potential Explanations

As for attempted explanations for the disciplinary differences, the review of the teaching evaluation questionnaire literature by Marsh (1987) originally treated disciplinary differences in ratings as a bias. Marsh (2007), though, does not include discipline in an extensive set of potential biases. Marsh (2007) uses Centra's (2003) definition of bias, 'Bias exists when a student, teacher, or course characteristic affects the evaluations made, either positively or negatively, but is unrelated to any criteria of good teaching, such as increased student learning' (p. 350).

This definition is readily applicable to other factors which have been discussed as potential biases, such as the size of classes and whether a course is taken as an elective major or compulsory minor. However, treating discipline as a bias is debateable. It does not seem reasonable to interpret discipline as having no relationship to good teaching. There is concurrence that course evaluation questionnaires give quite reliable and valid measures of teaching effectiveness (Aleamoni 1999; Feldman 1996, 1997, 2007; Marsh 1987, 2007; Marsh and Dunkin 1997; McKeachie 1997). If this position is accepted it implies that teaching tends to be more effective in some disciplines.

The original suggestion that disciplinary influence could be interpreted as a bias seems to have resulted in a limited discussion of an underlying reason for the differences in ratings. Neumann and Neumann (1985) had no evidence of underlying cause, but speculated on the reason their study found differences between 'hard' and 'soft' disciplines. They thought that soft subjects might have less agreement on knowledge development; so the classes would require discussion of major arguments. Hard subjects would be presented in a more routine format requiring a narrower range of teaching skills. However, if the necessary range of teaching skills is narrower, it might seem reasonable to assume that a higher proportion of teachers would have attained these necessary skills.

Feldman's (1997) review suggested a number of possible explanations, but it was clear that these potential explanations were purely speculations.

Among possible causes of these differences are the following: some courses are harder to teach than others; some fields have better teachers than others; and students in different major fields rate differently because of possible differences in their attitude, academic skills, goals, motivations, learning styles, and perceptions of good teaching. (p. 48)

Murray and Renaud (1995) investigated disciplinary variations in the frequency with which classroom teaching behaviours were utilised. They reported that teachers in the arts and humanities made greater use of behaviours classified as interaction, rapport and

mannerisms. Teachers in the natural sciences had a high incidence of organization and pacing behaviours. The study then proceeded to examine correlations between the frequencies of ten teacher behaviour dimensions with student ratings of teaching effectiveness for three discipline groups. Their inference was that the correlations were best interpreted as being random and hence indicating no significant differences between disciplinary groups. The conclusion was, therefore, that there were differences between the nature of teaching and disciplines manifest in differing frequencies of types of classroom behaviour. However, students' views of the effectiveness of the types of behaviour did not differ significantly between disciplines. This suggests that there is a common model of what constitutes good teaching which is independent of discipline.

This is consistent with Kember and McNaught (2007), who analysed interviews with 62 award-winning university teachers in Australia and Hong Kong. The teachers came from across the major discipline areas, yet it was possible to derive ten principles of good teaching which were consistent with the beliefs and practices of the whole sample. More stringent checks of consistency between teachers and disciplines were possible with a sub-sample of 18 Hong Kong award-winning teachers from all faculties of a comprehensive university (Kember et al. 2006). Transcripts were searched for utterances consistent with the principles and high levels of matching were found. Each of the 18 teachers was also asked to examine the conclusions for compatibility with their practices as a teacher and there was no disagreement.

The notion of a common inter-disciplinary model of good teaching is also consistent, implicitly at least, with the CTEQ literature. Major reviews of the field (Aleamoni 1999; Feldman 1996, 1997, 2007; Marsh 1987, 2007; Marsh and Dunkin 1997; McKeachie 1997) all discuss the nature of good university teaching, as the design of a valid instrument is conditional on identifying good teaching practice. There is recognition that there are alternative models of good teaching used to frame instruments, with the variation arising from differing theoretical perspectives of the originators. However, once developed from a model, it has been normal to use questionnaires across a broad range of disciplines. None of the reviews indicate that it is necessary to have instruments specific to particular disciplines and indeed there is a general advocacy of the common use of well-designed questionnaires with multi-factor structures corresponding to the identified facets of effective teaching.

It ought to be noted that there are contrary views to the idea of a common disciplinary view of what constitutes good teaching. Schulman (1986, 1987) advocated the concept of pedagogical content knowledge. Hativa and Marinovich (1995) interpret Schulman's work as implying that faculty developers who give counseling to academic staff about their CTEQ results, need to develop expertise with respect to discipline. While the notion of disciplinary differences in what constitutes good teaching might appeal intuitively to many academics, it is hard to find clear empirical evidence to support the claim.

It is possible that any differences between disciplines may be magnified by academics perceptions of them. Becher and Trowler (2001) argued that disciplines should be viewed from a perspective intermediate between realist and phenomenological positions. Their view of differences between disciplines was that there was an element of social construction. 'We need to take into account narratives, "stories", about disciplinary epistemology as well as disciplinary epistemology itself' (p. 38). Wareing (2009) could find little relevant evidence of differences between disciplines in how students learn. However, academics were prone to hold perceptions that good teaching and learning in their discipline differed from others.

Aims

The literature review clearly establishes that there have been consistent findings of disciplinary differences between student ratings of teaching. The Biglan (1973) classification scheme seems to provide a viable way of grouping the disciplines, with the hard v soft classification particularly apposite. There is evidence of commonality in students' perceptions of what constitutes good teaching across disciplines, but also contrary claims. It, therefore, seems particularly appropriate to test whether data on students' perceptions of the quality of the teaching and learning environment they experience fit to a common model across disciplines. If a common model were tenable it would then be apt to examine whether there were disciplinary effects for the magnitude of variables within the teaching and learning environment. It would be appropriate to include within a tested model a measure of perceptions of learning outcomes, so as to have a focus upon student learning rather than teacher performance. For all variables, student perceptions would be an adequate measure as the aim is to explain differences in student ratings by discipline. These are the perception measures widely used to appraise teaching quality and have provided the evidence of disciplinary differences; so it is appropriate to utilise them within a study of the phenomenon.

The aims of the study were formulated as:

1. To test whether a common model of good teaching applies across contrasting disciplinary groups.
2. To examine the extent of deployment between disciplines of variables within a broadly conceived teaching and learning environment.
3. To determine whether students perceive differing learning outcomes by discipline from any differences in the teaching and learning environment they experience.

Kember and his colleagues (Kember and Leung 2005a, b, 2006, 2009; Kember et al. 2007; Leung and Kember 2005a, b, 2006) had used structural equation modelling (SEM) to investigate the impact of a broadly defined teaching and learning environment upon the development of generic capabilities. The models tested in these studies incorporated a broadly conceived teaching and learning environment, including teaching and curriculum variables, together with variables pertinent to teacher–student and student–student relationships. It, therefore, encompassed the type of variables which Murray and Renaud's (1995) study indicated as being deployed to differing extents by discipline. The model hypothesised that the teaching and learning environment impacted upon the development of a set of generic capabilities. These capabilities were of the type that all college graduates are now expected to possess (Barrie 2006; Candy and Crebert 1991; Leckey and McGuigan 1997; Longworth and Davies 1996; Tait and Godfrey 1999), so this introduced an outcomes measure which could reasonably be expected to result from all disciplines.

The model hypothesised the nine scales of the teaching and learning environment influencing the development of the six capabilities. The teaching and learning environment scales acted as indicators for three latent constructs, while the capabilities were subsumed under two latent constructs. The model, shown in Fig. 1, aims to investigate the way capabilities could be nurtured through a teaching and learning environment of appropriate configuration and quality (Kember and Leung 2005a; Kember et al. 2007). The model, and a similar precursor one, had previously been tested with large multi-disciplinary samples in two universities (Kember and Leung 2005a, b, 2006, 2009; Kember et al. 2007; Leung and Kember 2005a, b, 2006). These studies did not specifically test for disciplinary differences, but the model showed a good fit to the data for samples with a wide range of disciplines, suggesting that the model is applicable across disciplines.

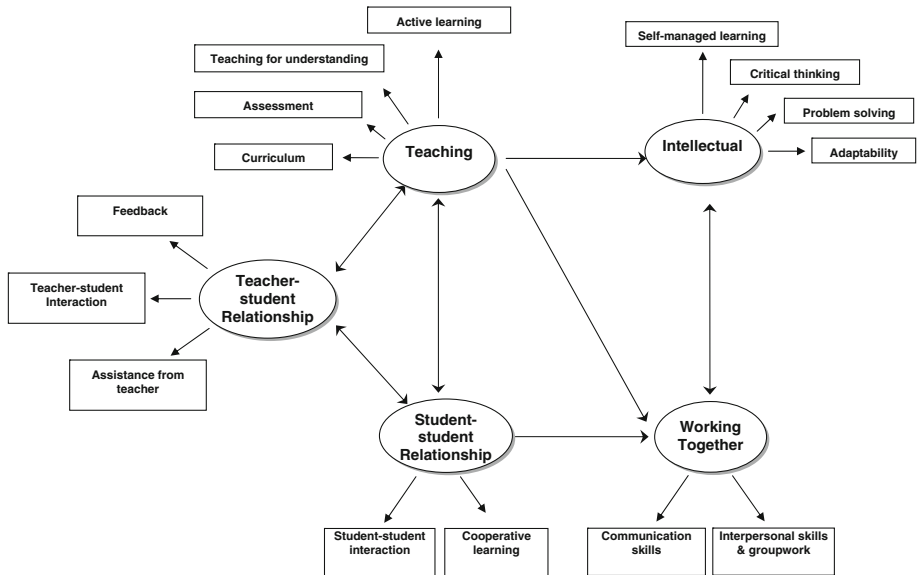


Fig. 1 The conceptual model relating the teaching and learning elements and the development of capabilities. *Note:* variances/disturbance terms of the latent constructs are not displayed for simplicity

The survey instrument was used to gather data from students in all undergraduate degrees within a comprehensive university in Hong Kong. The comprehensive university offered a wide range of disciplines, similar to that of other comprehensive research-intensive universities. It was, then, possible to compare models across broad groupings of disciplines, along traditional lines, using the group comparison techniques of SEM.

Four discipline areas were selected: arts, education and social science (labeled humanities); business administration (business); engineering and science (hard science); and health sciences and medicine (health). These disciplinary groupings were chosen so as to be consistent with the literature reviewed above and the Biglan (1973) category scheme. The hard versus soft distinction was the principal guide to discipline groupings as it had been most relevant in the CTEQ literature. Engineering and science were grouped together as initial courses in engineering are often used to build basic foundation knowledge, so are relatively pure in nature. The separate grouping for the health sciences was the most obvious manifestation of the life category. More fine-tuned groupings would have been undesirable as they would have complicated interpretation, made generalization more difficult and the smaller numbers in the resulting groups would have reduced the reliability of the SEM tests.

The humanities group included 11 departments or disciplines in arts and eight in social science, together with education. Business and administration offered programmes in business administration and finance. Hard science included students in 14 undergraduate degrees in science and from five engineering disciplines, with an emphasis on computer-based engineering. Health science was made up of medicine, nursing and pharmacy.

The aims given above can now be re-formulated in more specific terms, taking into account the nomination of the groups and the SEM techniques to be employed.

1. to test the model (Fig. 1) for configural invariance between the four groups
2. to compare latent and observed means for environment variables

- 3a. to compare latent and observed means for capability variables
- 3b. to compare the effects of the environment constructs on the capabilities.

Method

Sample and Procedures

Participants in the study were full-time undergraduate students from the 50 undergraduate programs offered by a university in Hong Kong. The universities in Hong Kong are governed by the University Grants Council (UGC), which has an international membership. All the universities were founded while Hong Kong was a British colony and are consistent with UK standards and practice. The UGC has been entrusted with ensuring that the standards and independence of Hong Kong universities has been maintained since the handover to China. Given the importance attached to globalisation and student exchange, the leading Hong Kong universities have become highly international in outlook, so can be seen as comparable to good quality universities elsewhere.

The questionnaires were administered to all 5,613 first- and third-year undergraduates and 3,341 of them completed and returned the questionnaire. Deletion of 36 cases with missing data ultimately yielded a final sample of size 3,305, 59% of the total population for the analysis. Due to the Hong Kong education system, most of the students were Chinese and aged between 18 and 22. The details of sampling procedure can be found in Kember and Leung (2006). The final sample was divided into the four discipline groups. The numbers within each group were; humanities ($n = 1,182$), business ($n = 694$); hard science ($n = 1,056$); health sciences ($n = 373$). The 50 undergraduate programmes were quite discrete. Apart from a small number of general education courses, most students would take courses associated with their major and few students would take courses from more than one of the disciplinary groupings.

Measures

The 33-item Student Engagement Questionnaire (SEQ) was used to seek feedback on the students' perceptions of the development of six generic capabilities and of nine elements in the teaching and learning environment (Kember and Leung 2009). The instrument and its precursor has been used extensively (Kember and Leung 2005a, b, 2006, 2009; Kember et al. 2007; Leung and Kember 2005a, b, 2006). Reliability, validity and other psychometric properties of the instrument have been dealt with in detail in Kember and Leung (2009).

The SEQ was designed to provide feedback at the level of a programme or degree. This distinguishes it from CTEQ-type instruments which focus at the level of individual teachers or courses. The most common equivalent is probably the Course Experience Questionnaire, originally developed by Ramsden and Entwistle (1981) and subsequently adapted for use nationally in Australia and the UK. The SEQ evaluates a broadly conceived teaching and learning environment (Kember and Leung 2009). As such it is responsive to diverse forms of teaching and learning, which appears to make it suitable for the present study.

Sample items for the instrument are presented in Table 1. Items were rated on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). In the following text,

Table 1 Student Engagement Questionnaire—sample items*Critical thinking*

Through this programme I have developed my ability to make value judgements about opposite perspectives

Self-managed learning

I have become more confident of my ability to pursue further learning

Adaptability

I have become more willing to change and accept new ideas

Problem solving

I am able to bring information and ideas together from different topics to solve problems

Communication skills

In this programme I have developed my ability to communicate effectively with others

Interpersonal skills and groupwork

I have learnt how to become an effective team or group member

Active learning

Students are given the chance to participate in class

Teaching for understanding

The teaching staff for this programme design classes with the aim of the students reaching an understanding of the course content

Feedback to assist learning

There was sufficient feedback on activities and assignments to ensure that we learnt from the work we did

Assessment

To do well in assessment in this programme you need to have good analytical skills

Teacher/student interaction

There is a close relationship between teaching staff and students

Assistance from teaching staff

I found teaching staff helpful when I had problems understanding the course content

Relationship with other students

My class groups have developed a strong sense of working together

Cooperative learning

I have frequently discussed ideas from courses with other students out-of-class

Coherence of curriculum

I can see how courses fitted together to make a coherent programme of study for my major

© 2003 David Kember, Doris Y.P. Leung and Carmel McNaught

we present the measures of internal consistency for each of the constructs, or latent variables, which were used in the structural equations models.

Intellectual

The intellectual capability was measured by eight items which were grouped into four scales with two items representing students' perceptions of the development of generic capabilities which are needed for a knowledge-based economic society (Candy and Crebert 1991; Leckey and McGuigan 1997; Longworth and Davies 1996; Tait and Godfrey 1999): (a) critical thinking ($\alpha = 0.78$), (b) self-managed learning ($\alpha = 0.74$), (c) adaptability ($\alpha = 0.61$), (d) problem solving ($\alpha = 0.70$).

Working Together

Four items were used to assess the development of students' capabilities in communication and team work skills. These items were grouped into two scales: communication skills ($\alpha = 0.72$), and interpersonal skills & groupwork ($\alpha = 0.57$).

Teaching

Students' perception of the teaching they received was assessed by ten items and were grouped into four constructs—that is, active learning ($\alpha = 0.70$), teaching for understanding ($\alpha = 0.81$), assessment ($\alpha = 0.57$), and coherence of curriculum ($\alpha = 0.79$). This measure was tapping both the teaching inside the classroom as well as the curriculum of the program.

Teacher–Student Relationship

Three constructs were measured by seven items to investigate students' perception of their relationship with their teaching staff—teacher/student interaction ($\alpha = 0.88$), assistance from teaching staff ($\alpha = 0.85$), and feedback to assist learning ($\alpha = 0.80$).

Student–Student Relationship

Student–student relationship was measured by four items comprising two constructs to reflect the perception of the bonding with peer students and engagement in learning activities among peer students, that is, relationship with other student ($\alpha = 0.86$), and cooperative learning ($\alpha = 0.71$).

The Cronbach alpha values for the scales range from 0.57 to 0.88; 13 of the 15 alphas exceed 0.60 and the other two are only marginally lower and in a range which has been argued to be acceptable (>0.5 ; Schmitt 1996).

Data Analysis

Data analyses were performed in a series of steps. First, we constructed summary scales based on the mean scores for items in a scale. Table 2 presents bivariate correlations for all of the 15 continuous variables in the hypothesized model for the overall sample ($n = 3,305$). The sizes of the correlations range from small (0.11) to moderate (0.65).

Next, we used SEM to test the research questions of our study using EQS 6.0 (Bentler 2006). First, data from each of the four discipline groups were separately tested to examine the degree of fit with the prior model (Fig. 1) and to demonstrate that the hypothesized model fitted the data from the four discipline groups simultaneously. Given the evidence for invariance across discipline groups, we next used multigroup SEM to compare factor loadings of our model across groups by constraining the factor loadings to be equal across groups while structural paths were allowed to be freely estimated (metric invariance). Then, observed mean-level differences among groups were compared by computing Cohen's d as a measure of effect sizes. Finally, we tested latent mean-level differences in the five constructs among the groups in two steps using mean and covariance structure analysis (MACS) which models both the pattern of means and covariations among the scales simultaneously (Bentler 2006; Byrne 2006). In the first step, we tested for intercept invariance by equating the intercepts of the 15 scales across groups to the final model in

Table 2 Bivariate correlations among the six scales in capabilities and nine scales in the teaching and learning environments of the overall sample ($n = 3,305$)

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Critical thinking	–														
2. Self-managed learning	0.26	–													
3. Adaptability	0.28	0.31	–												
4. Problem solving	0.40	0.37	0.41	–											
5. Communication skills	0.40	0.20	0.33	0.36	–										
6. Interpersonal skills & groupwork	0.22	0.21	0.38	0.30	0.43	–									
7. Active learning	0.38	0.16	0.19	0.24	0.33	0.22	–								
8. Teaching for understanding	0.27	0.17	0.16	0.21	0.19	0.14	0.49	–							
9. Feedback to assist learning	0.30	0.18	0.17	0.22	0.25	0.22	0.47	0.53	–						
10. Assessment	0.26	0.13	0.18	0.19	0.24	0.18	0.38	0.35	0.33	–					
11. Teacher–student interaction	0.29	0.14	0.16	0.20	0.25	0.20	0.47	0.45	0.55	0.32	–				
12. Assistance from teaching staff	0.26	0.21	0.18	0.23	0.22	0.20	0.39	0.45	0.65	0.29	0.55	–			
13. Relationship with other students	0.20	0.11	0.18	0.15	0.21	0.36	0.28	0.24	0.29	0.20	0.35	0.28	–		
14. Cooperative learning	0.25	0.22	0.21	0.24	0.18	0.23	0.27	0.24	0.25	0.25	0.28	0.26	0.42	–	
15. Coherence of curriculum	0.27	0.19	0.21	0.25	0.22	0.19	0.38	0.40	0.34	0.34	0.35	0.32	0.28	0.29	–

All p -values < 0.05

metric invariance. If intercept invariance was established, we further tested the scalar invariance. Latent means and mean differences for the five latent constructs were compared by selecting the humanities as the reference group and fixing latent means of this group at zero.

Assessment of overall model fit was based on multiple criteria including both absolute misfit and relative fit indices. The absolute misfit indices included the root mean square error of approximation (RMSEA; Browne and Cudeck 1993) and the standardized root mean squared residual (SRMR; Bentler 2006). Values of RMSEA and SRMR < 0.08 are indicative of an acceptable fit. The relative goodness-of-fit index was the comparative fit index (CFI; Bentler 1990). As a rule of thumb, values greater than 0.9 for CFI are considered as indicating an acceptable fit (Holye 1995), and values approaching 0.95 are indicative of a good fit (Hu and Bentler 1999). Models with both SRMR and CFI or both SRMR and RMSEA values indicating acceptable fit are not rejected.

The normality of all 15 scales was investigated for each of the four disciplinary groups and the distributions of all observed variables were found to be within the level recommended for SEM with maximum likelihood estimation procedure (skewness < 2 and kurtosis < 7) (West et al. 1995). Converged solutions with no out-of-range parameter estimates were obtained for all the analyses.

Results

Configural Invariance

The procedure involved initial SEM analyses to separately test the hypothesized model shown in Fig. 1 with data from the four discipline groups. Factor loadings for active learning, feedback to assist learning, relationship with other students, critical thinking, and communication skills were fixed to 1 for identification. The hypothesized model provides adequate approximations to the data from each of the four groups (Model 0 for each of the four groups) as indicated by the fit indexes in Table 3.

The next analysis was to test for configural invariance, which involved showing that the data from the four groups were consistent with the established model structure (Fig. 1). Multiple-group SEM was performed to assess for configural invariance across the four disciplinary groups by analysing the four samples simultaneously without imposing any constraints across the groups (Bentler 2006). The goodness-of-fit results for testing configural invariance are also shown in Table 2 (Model 1) which indicated an acceptable fit to the data ($\chi^2(332) = 1439.18$, $p < 0.001$, CFI = 0.92, RMSEA = 0.06, SRMR = 0.05). All indicators significantly loaded on their respective factors. Hence, it can be claimed that the structural forms of the models are the same for the four disciplinary groups. This implies a common mechanism for good teaching within the four disciplinary groups.

Metric Invariance

Given this preliminary evidence for invariance across discipline groups, we ran SEM to test metric invariance by further constraining factor loadings to be equal across groups (Model 2). The constrained model also showed acceptable fit, ($\chi^2(362) = 1531.04$, $p < 0.001$, CFI = 0.92, RMSEA = 0.06, SRMR = 0.06), however, the difference in chi-square between the constrained and unconstrained models was significant, $\Delta\chi^2(30) = 91.86$, $p < 0.001$, so full metric invariance was not supported.

Table 3 Results of testing invariance of the 5-factor conceptual model to the six scales in the capabilities and nine scales in the teaching and learning environment across disciplinary groups

	χ^2	df	CFI	SRMR	RMSEA (90% CI)
Model 0 _{Humanities} ($n = 1,182$)	499.64	83	0.925	0.045	0.065 (0.060, 0.071)
Model 0 _{Hard Sciences} ($n = 1,056$)	506.42	83	0.916	0.052	0.070 (0.064, 0.075)
Model 0 _{Business} ($n = 694$)	215.06	83	0.944	0.040	0.048 (0.040, 0.056)
Model 0 _{Health Sciences} ($n = 373$)	218.05	83	0.904	0.050	0.066 (0.055, 0.077)
Model 1: configural invariance (baseline model)	1439.18	332	0.923	0.047	0.064 (0.060, 0.067)
Model 2: metric invariance	1531.04	362	0.918	0.057	0.063 (0.059, 0.066)
Model 2a: partial metric invariance ^a	1465.05	357	0.923	0.049	0.061 (0.058, 0.065)
Model 3: Model 2a + equal intercepts	2524.48	402	0.923	0.060	0.064 (0.061–0.067)
Model 4: Model 2a + latent mean difference	1959.02	387	0.926	0.051	0.063 (0.059, 0.066)

CFI comparative fit index, SRMR standardized root mean squared residual, RMSEA (90% CI) root-mean-square error of approximation with 90% confidence interval

^a Model 2a was obtained by releasing five equality constraints in Model 2 including the factor loadings of understanding between the humanities, sciences, and health sciences groups; of assessment between the humanities and sciences groups; and of cooperative learning and curriculum between the sciences and business groups

We then tested for partial invariance (Byrne et al. 1989) by freeing the paths which significantly differed across groups using multivariate Lagrange multiplier (LM) tests (Stieger et al. 1985). A model in which five factor loadings of teaching for understanding (hard sciences vs health sciences, and hard sciences vs humanities), assessment (hard sciences vs humanities), cooperative learning (hard sciences vs business), and coherence of curriculum (hard sciences vs business) were allowed to vary, provided an adequate fit to the data ($\chi^2(357) = 1465.05$, $p < 0.001$, CFI = 0.92, RMSEA = 0.06, SRMR = 0.05), and in addition did not differ significantly from the unconstrained model ($\Delta\chi^2(25) = 25.87$, $p > 0.05$). This suggested that partial metric invariance held across the groups. Factor loadings of this final model for each discipline groups are shown in Fig. 2.

Observed and Latent Means Comparison

Means and standard deviations for capability variables and variables in the teaching and learning environment for each of the four discipline groups are given in Table 4. Table 5 permits comparison of the mean values by computing effect sizes, using Cohen's d as a measure across the four discipline groups. Cohen (1988) suggested Cohen's d (absolute value) below 0.2 be regarded as small, those between 0.2 and 0.5 medium and values over 0.5 be treated as a large effect size.

To have a clearer picture, we further assessed for disciplinary mean-level differences in latent constructs using multigroup MACS. The intercept invariance model fit the data adequately, $\chi^2(402) = 2524.48$, $p < 0.001$, CFI = 0.92, RMSEA = 0.06, SRMR = 0.06. The results suggested that students who have the same value on the construct would obtain the same value on the observed variable regardless of their group membership (Vandenberg and Lance 2000). We then proceeded to examine the mean-level differences among the five latent constructs across the four discipline groups. The latent mean difference model also provided a reasonable fit to the data as suggested by $\chi^2(387) = 1959.02$, $p < 0.001$, CFI = 0.93, RMSEA = 0.06, SRMR = 0.05. Latent mean difference

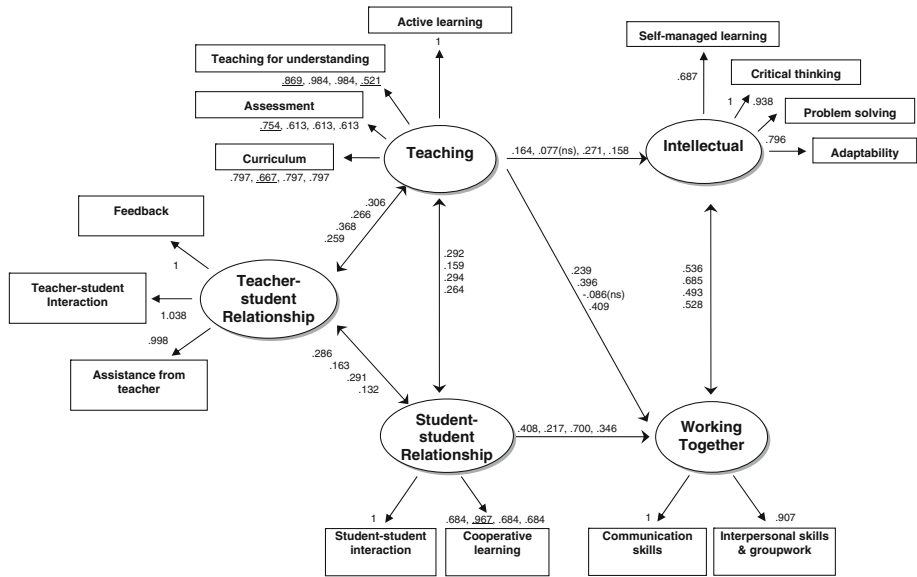


Fig. 2 Unstandardized solution (partial metric invariance) (humanities; business; hard science; health sciences). *Note:* variances/disturbance terms of the latent constructs are not displayed for simplicity. *Solution with one value* indicates the parameter estimate is invariant across the four groups, and with *value underlined* indicates the estimate is different from other groups

Table 4 Means and standard deviations of the six scales in capabilities and nine scales in the teaching and learning environment by disciplinary group

Scale	Humanities (n = 1,182)		Business (n = 694)		Hard sciences (n = 1,056)		Health sciences (n = 373)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Capability								
Critical thinking	3.72	0.82	3.40	0.82	3.26	0.90	3.47	0.83
Self-managed learning	3.97	0.70	4.06	0.68	3.98	0.76	4.12	0.63
Adaptability	3.85	0.66	3.92	0.66	3.86	0.70	3.86	0.64
Problem solving	3.69	0.69	3.75	0.67	3.72	0.73	3.79	0.68
Communication skills	3.50	0.90	3.64	0.87	3.00	0.92	3.68	0.89
Interpersonal skills & groupwork	3.34	0.86	3.57	0.79	3.33	0.85	3.51	0.86
Teaching and learning environment								
Active learning	3.09	0.88	2.86	0.85	2.84	0.90	3.37	0.86
Teaching for understanding	3.73	0.76	3.48	0.77	3.58	0.84	3.89	0.59
Feedback to assist learning	3.59	0.76	3.28	0.76	3.40	0.81	3.60	0.74
Assessment	3.54	0.78	3.49	0.73	3.43	0.78	3.67	0.76
Teacher–student interaction	3.49	0.92	3.23	0.86	3.26	0.96	3.57	0.86
Assistance from teaching staff	3.60	0.80	3.37	0.79	3.47	0.86	3.61	0.81
Relationship with other students	3.15	1.11	2.65	1.03	3.21	1.03	3.27	1.05
Cooperative learning	3.44	0.85	3.22	0.86	3.48	0.85	3.70	0.81
Coherence of curriculum	3.34	0.81	3.26	0.81	3.29	0.85	3.51	0.78

Table 5 Effect sizes (Cohen's *d*) of mean differences in the six scales in capabilities and nine scales in the teaching and learning environment across disciplinary group

	Humanities			Business		Hard sciences
	Business	Hard sciences	Health sciences	Hard sciences	Health sciences	Health sciences
Capability						
Critical thinking	0.39	0.53	0.30	0.16	-0.08	-0.24
Self-managed learning	-0.13	-0.01	-0.22	0.11	-0.09	-0.19
Adaptability	-0.11	-0.01	-0.02	0.09	0.09	0.00
Problem solving	-0.09	-0.04	-0.14	0.04	-0.06	-0.10
Communication skills	-0.16	0.55	-0.20	0.71	-0.05	-0.74
Interpersonal skills & groupwork	-0.27	0.01	-0.20	0.29	0.07	-0.21
Teaching and learning environment						
Active learning	0.26	0.28	-0.32	0.02	-0.59	-0.59
Teaching for understanding	0.33	0.19	-0.22	-0.12	-0.57	-0.39
Feedback to assist learning	0.41	0.24	-0.01	-0.15	-0.42	-0.25
Assessment	0.07	0.14	-0.17	0.08	-0.24	-0.31
Teacher–student interaction	0.29	0.24	-0.09	-0.03	-0.39	-0.33
Assistance from teaching staff	0.29	0.16	-0.01	-0.12	-0.30	-0.16
Relationship with other students	0.46	-0.06	-0.11	-0.54	-0.59	-0.06
Cooperative learning	0.26	-0.05	-0.31	-0.30	-0.56	-0.26
Coherence of curriculum	0.10	0.06	-0.21	-0.04	-0.31	-0.26

Table 6 Latent mean-level differences in the five latent factors in the conceptual model by disciplinary group

Latent construct	Hard sciences	Business	Health sciences	Humanities (reference)
Teaching	-0.191*	-0.232*	0.241*	–
Teacher/student relationship	-0.192*	-0.286*	0.012	–
Student/student relationship	0.043	-0.471*	0.209*	–
Intellectual	-0.230*	-0.144*	-0.112*	–
Working together	-0.345*	0.182*	0.179*	–

**p*-value < 0.003 indicates an overall type I error < 0.05 for the 15 constraints by Bonferroni corrections

estimates of the model are presented in Table 6. Bonferroni corrections were applied to account for the risk of capitalization on chance (Bollen 1989; Green and Babyak 1997) and the cut-off point of *p*-value < 0.003 (0.05/15 constraints) was used in assessing the significance of the latent mean differences. Compared with the humanities group, both the business and hard science groups exhibit significantly lower latent mean values in all the five latent constructs except the business group had a significantly higher mean value in working together and the hard sciences group had a non-significant difference mean value in student–student relationship; while the health sciences group scored significantly higher in teaching, student–student relationship, and working together, significantly lower in intellectual, and no difference in teacher–student relationship.

Comparing Effects of the Environment on Capability Development

The influence of the teaching and learning environment on capability development can, firstly, be examined by inspecting the models for each of the discipline groups with partial metric invariance (Model 2a). Figure 2 shows the unstandardized solution for each of the four groups by showing values for the unstandardized coefficients. Variances and disturbances to the latent variables and the values for them are not included in the diagram so as to make it conceptually simpler. The diagram includes parameters which are central to the research questions and the discussion which follows.

Table 7 enables comparison of the influence of the teaching and learning environment on capability development for the four discipline groups. The table gives direct, indirect and total effects for each of the five relevant paths in the model. The upper part of the Table gives non-standardized effects. These are normally used to examine the significance of effects. All of the factor loadings estimates of the models were statistically significant. However, the pattern of structural paths between the five latent constructs was different for the four groups in that the path from teaching to intellectual in the business group and the path from teaching to working together in the hard sciences group were statistically non-significant.

Discussion

Epistemological Beliefs

One of the frameworks drawn upon in this discussion is that of epistemological beliefs. The Biglan (1973) classification took into account epistemological differences in its derivation. It has subsequently been used to interpret epistemological issues and phenomena related to them. Becher and Trowler (2001) coined the notion of academic tribes to highlight differences between academic disciplines. They argued that values, ways of behaving and practices are related to the nature of knowledge and ideas in a discipline. Their book examined a range of aspects of academic life in terms of this framework, but paid relatively limited attention to teaching.

Smart and Ethington (1995) found differences by discipline in importance attached to goals for undergraduate education. Hard disciplines attached more importance to knowledge application, whereas soft ones thought knowledge integration was more important. Neumann and Neumann (1985) suggested that the better ratings of soft disciplines might occur because knowledge within them was less established and, therefore, more likely to be discussed in class. Since interaction and discussion had been associated with positive ratings (Murray and Renaud 1995) this suggested soft disciplines would tend to be rated better than hard ones.

Lindblom-Ylänne et al. (2006) studied variations in approaches to teaching by discipline. They used the Approaches to Teaching Inventory, which had two main scales; one for a conceptual change/student-focussed approach and the other for an information transmission/teacher-focussed approach. Teachers from hard disciplines were more likely to report a teacher-focussed approach, while those from soft disciplines made greater use of student-focussed approaches.

Humanities

In the models for the humanities and health groups, all paths between the teaching and learning environment and the capabilities latent variables were significant. These groups might, therefore, be regarded as conforming to the model in its purest form.

Table 7 Unstandardized and standardized total, direct and indirect effects of the three teaching and learning latent constructs on the two latent capability constructs by disciplinary group

	Humanities			Business			Hard sciences			Health sciences		
	Total effect	Direct effect	Indirect effect	Total effect	Direct effect	Indirect effect	Total effect	Direct effect	Indirect effect	Total effect	Direct effect	Indirect effect
Unstandardized structural path												
Teaching → intellectual	0.316*	0.163*	0.153*	0.375*	0.098	0.277*	0.247*	0.275*	-0.027	0.404*	0.153*	0.251*
Student–student relationship → intellectual	0.223*	-	0.223*	0.122*	-	0.122*	0.364*	-	0.364*	0.183*	-	0.183*
Working together → intellectual	0.544*	0.544*	-	0.643*	0.643*	-	0.502*	0.502*	-	0.551*	0.551*	-
Teaching → working together	0.281*	0.281*	-	0.430*	0.430*	-	-0.054	-0.054	-	0.456*	0.456*	-
Student–student relationship → working together	0.411*	0.411*	-	0.189*	0.189*	-	0.725*	0.725*	-	0.335*	0.335*	-
Standardized structural path												
Teaching → intellectual	0.360	0.186	0.174	0.441	0.115	0.326	0.275	0.305	-0.030	0.512	0.194	0.318
Student–student relationship → intellectual	0.320	-	0.320	0.184	-	0.184	0.435	-	0.435	0.254	-	0.254
Working together → intellectual	0.657	0.657	-	0.817	0.817	-	0.585	0.585	-	0.713	0.713	-
Teaching → working together	0.265	0.265	-	0.399	0.399	-	-0.051	-0.051	-	0.446	0.446	-
Student–student relationship → working together	0.487	0.487	-	0.226	0.226	-	0.743	0.743	-	0.357	0.357	-

**p*-value < 0.05

The humanities group had the highest latent mean for the intellectual capabilities and noticeably higher ratings for critical thinking than any of the other three groups. This is readily explainable in terms of the epistemological framework. The humanities group would be the one in which knowledge was most contested. Teaching was, therefore, likely to involve relatively high levels of discussion of alternative positions. This would particularly nurture critical thinking, but would also help in the development of the other intellectual capabilities (Kember and Leung 2005a, Kember et al. 2007). The interaction in class also tended to strengthen student–student relationships.

Health

The health group had intellectual capability latent means lower than for the humanities and comparable to business. All capability ratings were higher than for hard sciences. The ratings for the teaching and learning environment for health were the highest of the four groups. This data and the significance of all paths suggest no deviation from the pure model. The teaching and learning environment is optimised and as a consequence students perceive that the generic capabilities are nurtured.

The latent mean for the working together capabilities was comparable in value with business and higher than the other two groups. This is presumably because the teaching involves group activities giving practise in the working together capabilities, thus promoting their development as suggested by the highest standardised coefficient for the path between teaching and the working together capabilities among the four groups.

In terms of the epistemological framework, the health disciplines require students to acquire a body of basic knowledge which is reasonably well established. It is also important, though, that the necessary practitioner skills are also developed. These can best be developed through practise, so the health programs contain periods of clinical or professional practice as well as activities that simulate the practise of clinical skills. There are therefore significant portions of teaching devoted to active learning and interaction. There may also be an influence from these being caring professions. The teachers, most of whom would also be, or have been, practitioners in the field, may have developed an attitude of care towards their students and a heightened ability to counsel them. This would further contribute to the degree of teacher–student interaction.

Hard Science

The hard science disciplinary grouping has the highest standardised coefficient for the direct link from the teaching latent variable to the intellectual capabilities. The reason for this is possibly indicated by the high loading of teaching for understanding on the teaching latent variable. Presumably students perceive a high importance being attached to ensuring that they have a good understanding of important constructs. Science has a foundation body of well-established knowledge that students need to understand as a prerequisite for learning more advanced concepts.

The direct path from teaching to the working together capability latent variable was non-significant for hard science, which indicates that the method of teaching did not encourage the development of the working together capabilities. This might be because the hard science teachers concentrate upon instilling knowledge of well-established concepts through a predominantly didactic form of teaching. This provides limited opportunities for students to work together in groups or engage in discussion and hence limited opportunities to practise and develop the working together capabilities.

Indications of the didactic nature of the teaching come from the low values for the latent means for teaching and teacher–student relationships. Teachers in the disciplines are not effectively making use of the range of teaching behaviours and learning activities encompassed within the broadly conceived teaching and learning environment. The low latent mean for teacher–student relationships suggests a limited degree of interaction because of didactic teaching.

For the hard science group, the standardised coefficient for the direct path from student–student relationships to working together was particularly high at 0.74. This path coefficient was comparatively high for the indirect path from teaching, through working together to intellectual capabilities. Presumably this path is too indirect, though, to compensate for the non-significant teaching to working together path. The outcome is a negligible indirect effect of teaching on the development of intellectual capabilities in hard science.

While the direct effect of teaching on intellectual capabilities for hard science was the highest of the four groups, the total effect ends up as the least because of the non-contribution of indirect effects. This might go some way to explaining the relatively negative student perceptions of hard science teaching in the literature.

The didactic teaching is presumably adopted because teachers in the disciplines feel that, in the early years of the degree particularly, they are teaching a well-established body of knowledge. They conceive their role as teachers to transmit this body of knowledge to their students. As the knowledge is well established the need for discussion and active learning experiences might not seem apparent.

Hard vs Soft

In relating the findings of this study to the exiting literature, it is the comparison of hard and soft disciplines which is most pertinent. Most studies which offer any insights on why there are disciplinary differences conceptualise the issue as a hard/soft spectrum, so there is little to compare to the results for the health and business groupings of this study.

Epistemological differences have been used as an explanation for disparities in ratings and approaches to teaching between science and the humanities. Smart and Ethington (1995) found that, compared to hard disciplines, academics in soft disciplines placed more emphasis on knowledge acquisition and integration and less on application. This presumably links to the finding that teachers in soft disciplines are more likely to employ active learning methods (Braxton et al. 1988; Lattuca and Stark 1995).

However, academics in hard disciplines placed more value in undergraduate research projects (Lattuca and Stark 1995). The research behaviour of scientists and engineers has also been characterised by relatively high numbers of publications compared to their counterparts in the arts and social sciences (Becher and Trowler 2001).

The comparisons seem somewhat dichotomous. The teaching behaviour of academics in science appears to be more didactic because of claims that knowledge in the disciplines is more certain. However, undergraduate research is more valued and, if publications are a reliable indicator, there is more research activity. This might give credence to the idea that disciplinary behaviours are formed as much by socially-constructed stories as by real differences (Becher and Trowler 2001; Wareing 2009). The observation could be particularly applicable to teaching behaviour. Requirements for training in teaching for university teachers are limited and often non-existent. The major influence on the teaching of new academics is, then, the behaviour of their former teachers. Teachers in hard disciplines have been found to be more likely to have teacher-centred conceptions of teaching

(Lindblom-Ylänne et al. 2006), which means that, as students, new academics will have had considerable exposure to didactic teaching.

This suggested explanation attests to the influence of academic tribes (Becher and Trowler 2001). Disciplinary tribes construct their own culture of teaching, influenced by real and perceived epistemological differences. These are passed on to succeeding generations of the tribe through exposure to beliefs and resulting practices during their education.

Business Administration

For the business administration discipline, the latent mean for the working together capabilities is the highest of the four groups. The reason seems to be connected with the nature of teaching and learning, since the standardised coefficient from teaching to the working together capabilities is the highest of the four groups. This might be interpreted as meaning that the business administration discipline set a relatively high proportion of learning activities, such as group projects, which involve students working together. There might also be requirements for students to make presentations, which would help the development of communication skills. The programs also organised extra-curricular activities, such as visits and competitions, which help nurture working together qualities. This also provides a reason for non-invariance of the factor loadings of the student–student relationship with the other two disciplinary groups.

The business administration group had the lowest latent mean for teaching. The loadings on assessment and coherence of the curriculum were the lowest for the four groups, which presumably indicates students' perceptions of these were less strong. The latent means for intellectual capability development for the business administration and other groups were not significantly different. The lower teaching latent mean meant a smaller extent in the development of intellectual capabilities via direct impact. The indirect contribution for business was higher, though, because of the stronger impact of teaching upon the working together capabilities.

The teaching in business therefore placed an emphasis on developing working together capabilities through learning activities. There was less stress than science on teaching a body of knowledge. This is presumably because the disciplinary knowledge is not as well-established, particularly in management and marketing, while greater importance is attached to the development of business skills.

Conclusion

There has been an abundant literature attesting to variations in students' ratings of teaching by discipline and the patterns of variation have been quite consistent over numerous studies. However, there have been limited insights into why these occurred, and few of the attempted explanation have been backed with empirical evidence. The basis of using CTEQs as quality assurance measures is that they give reasonably reliable and valid measures of teaching quality. This suggests that students commonly report perceptions of variations in quality in teaching which show quite systematic distinction by discipline.

In this study we established that data from four disciplinary groups fitted to a common model of good teaching influencing the development of generic capabilities. Use of multiple-group SEM also showed configural invariance which reinforced the conclusion that the nature of an effective teaching and learning environment was consistent between

disciplines. There were, however, differences in the magnitude of structural paths and latent means. This implies that there were differences between disciplines in the extent to which elements within the teaching and learning environment were brought into play. It was possible to suggest reasons for these disciplinary variations in terms of the epistemological nature of the disciplines. Though, it appears that socially-constructed stories might play as much a part in epistemological beliefs as real disciplinary differences.

The study was confined to one university, so generalization is debatable. The patterns of variation in ratings within the sample were entirely consistent with those found in large studies and meta-analyses elsewhere. There is also prior evidence to back the inference of a common model of good teaching across disciplines, but differing degrees of implementation. Interpreting disciplinary differences in terms of epistemological distinctions has been common for other phenomenon. It is, therefore, plausible that the suggested explanation could also apply in other contexts. This is particularly the case if there is acceptance of the suggested influence of disciplinary tribes, in influencing the formation of socially-constructed beliefs about epistemology. Disciplinary tribes are international, so their influence should also be international.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Aleamoni, L. M. (1999). Student rating myths versus research facts from 1924 to 1998. *Journal of Personnel Evaluation in Education*, 13(2), 153–166.
- Barnes, M. W., & Patterson, R. H. (1988). Using teaching evaluation results to plan departmental strategies to accomplish the institutional mission. Paper presented at the annual meeting of the Society for College and University Planning, Toronto, ON.
- Barrie, S. (2006). Understanding what we mean by the generic attributes of graduates. *Higher Education*, 51(2), 215–241.
- Becher, T., & Trowler, P. R. (2001). *Academic tribes and territories: Intellectual enquiry and the cultures of disciplines*. Milton Keynes, UK: SRHE and Open University Press.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238–246.
- Bentler, P. M. (2006). *EQS structural equations program manual*. Encino, CA: Multivariate Software.
- Biggs, J. (1987). *Student approaches to learning and studying*. Melbourne: Australian Council for Educational Research.
- Biglan, A. (1973). The characteristics of subject matter in different academic areas. *Journal of Applied Psychology*, 57, 195–203.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Braxton, J. M., Olsen, D., & Simmons, A. (1988). Affinity disciplines and the use of principles of good practice for undergraduate education. *Research in Higher Education*, 39(3), 299–318.
- 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. 136–161). Newbury Park, CA: Sage.
- Byrne, B. M. (2006). *Structural equation modelling with EQS: Basic concepts, applications, and programming* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105, 456–466.
- Candy, P. C., & Crebert, R. G. (1991). Lifelong learning: An enduring mandate for higher education. *Higher Education Research and Development*, 10, 3–18.
- Cashin, W. E. (1990). Students do rate different academic fields differently. *New directions for teaching and learning*, 43, 113–121.
- Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education*, 44(5), 495–518.

- Cohen, J. (1988). *Statistical power analysis for the behavioural science* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Entwistle, N. J., & Ramsden, P. (1983). *Understanding student learning*. London: Croom Helm.
- Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers; what we know and what we don't. *Research in Higher Education*, 9, 199–242.
- Feldman, K. A. (1996). Identifying exemplary teaching: Using data from course and teacher evaluations. *New Directions for Teaching and Learning*, 65, 41–50.
- Feldman, K. A. (1997). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 368–395). New York: Agathon Press.
- Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 93–126). New York: Springer.
- Green, S. B., & Babyak, M. K. (1997). Control of Type I errors with multiple tests of constraints in structural equation modelling. *Multivariate Behavioral Research*, 32, 39–51.
- Hativa, N., & Marincovich, M. (1995). *Disciplinary differences in teaching and learning: Implications for practice*. San Francisco: Jossey-Bass.
- Holye, R. H. (1995). *Structural equation modelling: Concepts, issues, and applications*. Thousand Oaks, CA: Sage.
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55.
- Kember, D., & Leung, D. Y. P. (2005a). The impact of the teaching and learning environment on the development of generic capabilities needed for a knowledge-based society. *Learning Environments Research*, 8, 245–266.
- Kember, D., & Leung, D. Y. P. (2005b). The influence of active learning experiences on the development of graduate capabilities. *Studies in Higher Education*, 30, 155–170.
- Kember, D., & Leung, D. Y. P. (2006). Characterising a teaching and learning environment conducive to making demands on students while not making their workload excessive. *Studies in Higher Education*, 31, 185–198.
- Kember, D., & Leung, D. Y. P. (2009). Development of a questionnaire for assessing students' perceptions of the teaching and learning environment and its use in quality assurance. *Learning Environments Research*, 12, 15–29.
- Kember, D., Leung, D. Y. P., & Ma, R. S. F. (2007). Characterising learning environments capable of nurturing generic capabilities in higher education. *Research in Higher Education*, 48, 609–632.
- Kember, D., Ma, R. S. F., McNaught, C., & 18 excellent teachers. (2006). *Excellent university teaching*. Hong Kong: Chinese University Press.
- Kember, D., & McNaught, C. (2007). *Enhancing university teaching: Lessons from research into award winning teachers*. Abingdon, Oxfordshire: Routledge.
- Kwan, K. P. (1999). How fair are student ratings in assessing the teaching performance of university teachers? *Assessment and Evaluation in Higher Education*, 24, 181–195.
- Lattuca, L. R., & Stark, J. S. (1995). Modifying the major: Discretionary thoughts from ten disciplines. *Review of Higher Education*, 18(2), 315–344.
- Leckey, J. F., & McGuigan, M. A. (1997). Right tracks—wrong rails: The development of generic skills in higher education. *Research in Higher Education*, 38, 365–378.
- Leung, D. Y. P., & Kember, D. (2005a). Comparability of data gathered from evaluation questionnaires on paper and through the internet. *Research in Higher Education*, 46, 571–591.
- Leung, D. Y. P., & Kember, D. (2005b). The influence of the part-time study experience on the development of generic capabilities. *Journal of Further and Higher Education*, 29, 91–101.
- Leung, D. Y. P., & Kember, D. (2006). The influence of teaching approach and teacher–student interaction on the development of graduate capabilities. *Structural Equation Modeling*, 13, 264–286.
- Lindblom-Ylänne, S., Trigwell, K., Nevgi, A., & Ashwin, P. (2006). How approaches to teaching are affected by discipline and teaching context. *Studies in Higher Education*, 31, 285–298.
- Longworth, N., & Davies, W. K. (1996). *Lifelong learning*. London: Kogan Page.
- 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.
- 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., & Dunkin, M. J. (1997). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 241–320). New York: Agathon Press.
- McKeachie, W. (1997). Student ratings: The validity of use. *American Psychologist*, *52*, 1218–1225.
- Murray, H. G., & Renaud, R. D. (1995). Disciplinary differences in classroom teaching behaviours. *New directions for teaching and learning*, *64*, 31–39.
- Neumann, L., & Neumann, Y. (1985). Determinants of students' instructional evaluation: A comparison of four levels of academic areas. *Journal of Educational Research*, *78*, 152–158.
- Ramsden, P., & Entwistle, N. J. (1981). Effects of academic departments on students' approaches to studying. *British Journal of Educational Psychology*, *51*, 368–383.
- Schmitt, M. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350–353.
- Schön, D. A. (1987). *Educating the reflective practitioner*. San Francisco: Jossey-Bass.
- Schulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, *15*, 4–14.
- Schultman, L. S. (1987). Knowledge and teaching foundations of the new reform. *Harvard Educational Review*, *57*, 1–22.
- Smart, J. C., & Ethington, C. A. (1995). Disciplinary and institutional differences in undergraduate education goals. In N. Hativa & M. Marincovich (Eds.), *Disciplinary differences in teaching and learning: Implications for practice* (pp. 40–58). San Francisco: Jossey-Bass.
- Stieger, J. H., Shapiro, A., & Browne, M. W. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika*, *50*, 253–264.
- Tait, H., & Godfrey, H. (1999). Defining and assessing competence in generic skills. *Quality in Higher Education*, *5*, 245–253.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–70.
- Wareing, S. (2009). Disciplines, discourse and Orientalism: The implications for postgraduate certificates in learning and teaching in higher education. *Studies in Higher Education*, *34*(8), 917–928.
- West, S. G., Finch, J. F., & Curren, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modelling: Concepts, issues, and applications* (pp. 56–75). Thousand Oaks, CA: Sage.