Educational Assessment, Evaluation and Accountability (2020) 32:481–498
https://doi.org/10.1007/s11092-020-09336-6

Online vs. offline course evaluation revisited: testing the invariance of a course evaluation questionnaire using a multigroup confirmatory factor analysis framework

Ellen Laupper[1] · Lars Balzer[1] · Jean-Louis Berger[2]

Received: 30 July 2019 / Accepted: 5 October 2020 / Published online: 6 November 2020

Abstract
Survey-based formats of assessing teaching quality in higher education are widely used and will likely continue to be used by higher education institutions around the world as various global trends contributing to their widespread use further evolve. Although the use of mobile devices for course evaluation continues to grow, there remain some unresolved aspects of the classic paper and web-based modes of evaluation. In the current study, the multigroup confirmatory factor analysis approach (MGCFA), an accepted methodological approach in general mixed-method survey research, was chosen to address some of the methodological issues when comparing these two evaluation modes. By randomly assigning one of the two modes to 33 continuing training courses at a Swiss higher education institution, this study tested whether the two different modes of assessing teaching quality yield the same results. The practical implications for course evaluation practice in institutions of higher education as well as the implications and limitations of the chosen methodological approach are discussed.

Keywords Course evaluation · Web survey · Multigroup confirmatory factor analysis · Invariance measurement · Higher education · Mode effect

✉ Ellen Laupper
ellen.laupper@sfivet.swiss; https://www.ehb.swiss/

Lars Balzer
lars.balzer@sfivet.swiss

Jean-Louis Berger
jean-louis.berger@unifr.ch

[1] Evaluation Unit, SFIVET Swiss Federal Institute for Vocational Education and Training, Kirchlindachstrasse 79, Postfach, CH-3052 Zollikofen, Switzerland

[2] Department of Education, University of Fribourg, Rue P.A. de Faucigny 2, CH-1700 Fribourg, Switzerland

1 Introduction

Originally, assessment of teaching quality was primarily used for formative purposes. However, during the 1970s in the USA, its use started to expand to human resources decisions concerning faculty personnel. Then, at the beginning of the new millennium, in the course of increased international cooperation and competition, a process of legalising quality for higher education institutions started, which created a need for practices of accountability (Borch et al. 2020; Donzallaz 2010; Skedsmo 2020; Spooren et al. 2017).

Despite the well-known critiques of the use of survey-based formats to assess teaching quality in higher education, higher education institutions all over the world continue to use them (Spooren et al. 2017). This trend is fuelled by the abovementioned changes in the intentions of the use of such information over the last half century as well as the relative ease of implementing and standardising the procedure and collecting, processing, and communicating large amounts of data. To date the often simultaneous use of the method for different purposes has been critically discussed by practitioners and researchers alike, and various alternative or supplementary approaches have been tested, suggested, and implemented to address different stakeholders' need for different information (e.g. Borch et al. 2020; Skedsmo 2020; Spooren et al. 2017). However, as the institutional accreditation of higher education institutions will continue to be a vital requirement and, as mentioned above, this survey method produces comparable results with relative ease, higher education institutions will likely continue to rely on this method. Therefore, it will remain important to ensure that this method produces valid and reliable data.

This is all the more important since online evaluations of teaching in higher education have become increasingly common since the beginning of the new millennium (Crews and Curtis 2011; Morrison 2011; Venette et al. 2010; Treischl and Wolbring 2017). The method's practicality, feasibility, flexibility, time- and cost-effectiveness when dealing with large samples and a large amount of data, and its potential to provide real-time feedback make it an attractive option that is increasingly replacing paper-and-pencil course evaluations (Barkhi and Williams 2010; Dommeyer et al. 2004; Hessius and Johansson 2015; Layne et al. 1999; Nulty 2008; Risquez et al. 2015). As nowadays mobile devices, such as smartphones or tablet computers, are more and more widely used on campuses all over the world, a new line of discussion moves toward the practicability and feasibility of 'mobile' course evaluations (Champagne 2013; Hessius and Johansson 2015). This raises the question whether we must consider research on issues related to the data quality of paper-and-pencil and online course evaluations—that is, web surveys completed on a desktop or laptop computer—to be part of a 'historical' (Champagne 2013, p. 644) debate that is no longer relevant.

2 Current data quality issues in comparative research on the effects of the online and offline survey mode

As the use of mobile devices grows, a new data-gathering survey mode (survey method) for course evaluation is emerging. However, there are still some aspects of the differences between the two classic modes of course evaluation (i.e. paper-and-pencil and online) that must be discussed, as recent studies addressing this issue face some essential methodological problems (Capa-Aydin 2016; Mitchell and Morales 2017; Morrison 2013; Risquez et al. 2015). Because of the methodological and theoretical shortcomings of the 'old' online/offline mode difference studies and the lasting controversial discussion concerning the usefulness of course evaluation, especially student evaluation of teaching (SET) (e.g. Rienties 2014), the data quality of classic online course evaluation is still a topic of interest. Therefore, the paper-and-pencil vs. online debate should continue.

Manifold quality indices have been studied in previous online/offline mode research, including indices of missing data, such as item nonresponse, reliability, biases in reply patterns or the number of words written, and the richness of information provided in answers to open-text questions (e.g. Deutskens et al. 2006; Hardré et al. 2012). A large body of research focuses on comparing response rates and overall or item means (for an overview, see Capa-Aydin 2016; Mitchell and Morales 2017; Morrison 2013). While small and inconsistent (or no) mode differences have been observed for most of the compared indices (Deutskens et al. 2006), consistently lower response rates for online course evaluations than for paper course evaluations have been reported (e.g. Klieger et al. 2014; Shih and Fan 2008). It is also widely agreed that online course evaluations provide more and richer information for open-text questions (Deutskens et al. 2006; Kays et al. 2012; De Leeuw and Hox 2011). Usually, no difference or lower item means have been found for online evaluations, although some studies have found slightly higher mean ratings for this mode (for overviews, see Capa-Aydin 2016; Mitchell and Morales 2017; Morrison 2013).

Given these inconclusive findings (see, e.g. Klieger et al. 2014) and the fact that in-class paper-and-pencil surveys are most often compared to out-of-class online surveys (Capa-Aydin 2016), one line of research is looking for confounding psychological factors of the completion situation affecting data quality to explain some of the findings.

These factors include, for example, (pseudo)voluntariness, perceived anonymity, social desirability, and cognitive load (Dittmann-Domenichini and Halbherr 2015; Hardré et al. 2010; Hardré et al. 2012; Kordts-Freudinger and Geithner 2013).

While this line of research certainly has the potential to provide in-depth knowledge on how mode effects are mediated by psychological factors, many of the studies, including some of the most recent ones, still suffer from important methodological limitations.

As this research is often conducted by in-house evaluation units that use the window of opportunity when faculties or universities decide to switch the mode of course evaluation, it is unsurprising that the study designs, range of study populations, and instruments vary considerably (Mitchell and Morales 2017; Spooren et al. 2013). Despite this heterogeneity, several critical issues previously raised by researchers, which considerably influence the validity of the study results, could be avoided by including some methodological considerations (e.g. Capa-Aydin 2016; Klieger et al. 2014; Morrison 2013; Spooren et al. 2013; Stowell et al. 2012). One of the main methodological problems that has been raised is the lack of control for possible confounding factors by randomisation, matching or statistical methods. This is problematic for several reasons. First, the response rates for online evaluations are substantially lower, and it is well known that mode preferences are highly biased (e.g. Avery et al. 2006). Second, research has shown that course evaluation ratings are influenced by a broad variety of student, teacher, and course characteristics (for an overview, see Spooren et al. 2013). Thus, these data are of a nested nature, at least on the course level (Risquez et al. 2015; Morrison 2013). It is therefore difficult to draw generalisable conclusions about mode effects from such studies.

Morrison (2013) considered 24 'key studies', of which only 5 explicitly reported randomisation at the student level. In her overview of 15 studies, Capa-Aydin (2016) pointed out that only 9 of 12 studies with an experimental design reported to have conducted randomisation. She also found that only a few studies validated their instrument before using it for comparison, and some used statistical tests inadequately by conducting multiple comparisons at the item level without correcting for an increase in type I error. Thus, mode differences may have been over-interpreted by previous studies.

As Spooren et al. (2013) summarised in their discussion of the state-of-the-art knowledge about SET, the dimensionality debate remains unresolved. Although it is widely accepted that SET instruments should be multidimensional to capture multiple aspects of the quality of teaching, a strong position still tends to favour a single overall score. These unresolved issues are rooted in the multiple functions that SET instruments often have to serve (see also Spooren et al. 2013). This seems to have an effect on SET mode research, as most studies focus mainly on comparing single-item or overall means instead of the comparability of the instrument as a whole. The whole picture of possible mode effects cannot be captured if only mean differences are considered, as these are only an indicator of a shift in the response distribution. Moreover, they are often not effective for distinguishing between systematic and random measurement errors. Only comparison of the factor structure and relationships between the dimensions can reveal whether the whole construct of teaching quality and its various dimensions are measured in the same way by the two modes. A difference in the factor structure would imply that mode differences, such as differences in visual presentation, data input by clicking as opposed to writing, or the presence or absence of the lecturer, would influence the respondents' evaluative process (Hox et al. 2015). To our knowledge, only a few studies have compared the factor structure of their SET instruments between modes (i.e. Capa-Aydin 2016; Morrison 2013; Layne et al. 1999; Leung and Kember 2005). In addition, most have used explorative approaches. However, an explorative approach is more suitable during the instrument development process, when the factor structure of an instrument is not yet known. Furthermore, with this approach, the comparability of two different factor structures can only be assessed in a descriptive way, as no straightforward statistical procedure exists. In contrast, with confirmatory factor analysis (CFA), a factor structure testing procedure exists, which allows one to statistically test the fit of a theoretically assumed factor structure and assess the relationships between the items and factors as well. Furthermore, it offers the possibility to statistically test the equality of a theoretically assumed factor structure across two or more subgroups simultaneously (see Romppel 2014; Sass and Schmitt 2013).

In the broader scientific field of survey methodology research, which has extensively examined systematic control of measurement errors, mode effects are also a topic of interest (e.g. Biemer et al. 2017). Especially in research on mixed-mode survey designs, which use paper-and-pencil and online questionnaires—among other modes—

simultaneously or consecutively in the same study, multigroup CFA has proven to be feasible and valuable for testing the comparability of data obtained with different modes (De Leeuw and Hox 2011; Gregorich 2006; Hox et al. 2015; Klausch et al. 2013; Schmitt and Kuljanin 2008; Vandenberg and Lance 2000). In this paper, by using the multigroup CFA strategy, we would like to address some of the methodological issues identified in previous research on SET mode differences.

3 Testing the invariance of a questionnaire with a multigroup confirmatory factor analysis framework

Structural equation modelling (SEM) is a statistical methodology for modelling the latent structure of an underlying set of observed variables. SEM can statistically test the model fit of the data for all assumed relationships simultaneously. Although multigroup invariance testing based on confirmatory factor analysis (MGCFA) is mainly used to assess the invariance of an instrument between gender, racial, cultural, linguistic, or other sociodemographic diverse subgroups, several studies have used this method to assess the invariance between different modes of survey administration (Schmitt and Kuljanin 2008). Since this MGCFA testing procedure is not broadly known in the SET literature, the different forms of measurement invariance and their meaning are briefly described here in nontechnical language. Furthermore, since there are several publications on this method, which differ slightly in their recommendations regarding the consecutive steps and the terms for the different invariance concepts (e.g. Hox et al. 2015; Sass and Schmitt 2013; Steenkamp and Baumgartner 1998; van de Schoot et al. 2012), we decided to follow the in detail outlined test strategy proposed by Byrne (2012). We also referred to other sources, mostly Steinmetz et al. (2009) and Gregorich (2006).

According to Steinmetz et al. (2009), the following four questions can be addressed with MGCFA: (1) Are the measurement parameters (factor loadings, measurement errors, etc.) the same across groups? (2) Are there pronounced response biases in a particular group? (3) Can one unambiguously interpret observed mean differences as latent mean differences? (4) Is the same construct measured in all groups? The testing procedure proposes steps that build upon one another, forming a nested hierarchy of progressively more restricted invariance hypotheses by constraining more and more measurement parameters equal in order to allow different conclusions about the acceptable level of comparability.

After testing the fit of the baseline model (i.e. the postulated structure of the measurement instrument under study) for each group separately, the first step of MGCFA is testing for dimensional and configural invariance for both groups in a combined multigroup model without imposing any equality constraints (Gregorich 2006). When confirmed, these two invariance levels only imply that the instrument has the same number of factors and that each factor is associated with an identical item set across groups (Gregorich 2006).

To establish metric invariance, the factor loadings must be equal across groups. This level of invariance indicates that the respondents attribute the same meaning to the factors of a construct across groups (van de Schoot et al. 2012). If the metric invariance hypothesis is not supported by the data, two possible interpretations should be considered. Either one or more factors—or a subset of items—may have a different meaning across groups, or a subset of factor loading estimates for one group may be biased by a response style. That is, respondents either tend to use systematically extreme response options, such as ‘never’ or ‘always’, or systematically avoid extreme response options. Both response styles affect response variation (Gregorich 2006).

Another threat to measurement invariance is the differential additive response bias or differential acquiescence response style (Gregorich 2006). Forces unrelated to the factors, such as different cultural or gender norms in the groups, or, in this case, maybe forces related to administration mode differences could cause items to be systematically higher or lower valued. To rule out such a bias and thus establish scalar invariance (Steenkamp and Baumgartner 1998; Steinmetz et al. 2009; van de Schoot et al. 2012), the item intercepts must be equal across the groups. In contrast to the response styles mentioned above, this bias does not affect response variation; instead, it relates to the value of the observed means. When not equally present in all compared groups, this bias contaminates the estimates of group mean differences, thereby threatening the assumption that differences in the means of the observed items are related to differences in the means of the underlying construct(s) (Steenkamp and Baumgartner 1998). Therefore, only if scalar invariance is established a comparison of the groups latent means is meaningful (van de Schoot et al. 2012).

To further assess whether the factorial structure is invariant across groups, the factor variances, covariances, and error variances must be equal (Byrne 2012). Thereby, the invariance of factor variances is a prerequisite for interpreting equal factor covariances as correlations and equal error variances as equal reliabilities (Steinmetz et al. 2009). However, as Bialosiewicz (2013, p. 9) notes, ‘…strict invariance [which the author defines as having two sublevels: (1) invariance of factor variances and (2) invariance of error variances] represents a highly constrained model and is rarely achieved in practice’. This is why achieving this level of invariance across groups is seen as an unrealistic standard by most experts in the field. Furthermore, for comparison of group means, it is of limited practical value (Gregorich 2006).

3.1 Research questions and hypothesis

The overall aim of this study was to test whether it is defensible to replace an already used paper-and-pencil course evaluation questionnaire with a corresponding online course evaluation questionnaire. More specifically, we were interested in the question of whether the two administration modes produce differences in the survey data. In other words, are course evaluation results collected with a paper-and-pencil questionnaire comparable to the results collected with an online questionnaire? This was tested with a MGCFA following the methodological procedure described above. The following research questions were posed:

- Does the factor structure for both modes match the proposed theoretical model of the course evaluation questionnaire? (dimensional and configural invariance)
- Are the factors measured with the same accuracy across modes? (metric and scalar invariance)
- Are the relationships between the latent factors the same across modes? (invariance of factor variances and factor covariances)

- Are the factor means for the two administration mode groups the same? (invariance of latent means)

Based on the non-existent or limited and inconsistent findings of comparisons of item means in SET research so far (see discussion on p. 4), we assume that the administration mode of our course evaluation questionnaire does not affect the resulting course evaluation. We therefore expect no differences in the various invariance parameters (measurement errors, factor loadings, intercepts, factor variances, factor covariance, and latent means) when comparing the course evaluation results collected with a paper-and-pencil questionnaire to the results collected with a corresponding online questionnaire.

4 Methodology

4.1 Design and procedure

In the pilot phase, which led to the development of a programme for evaluating continuing training courses for vocational and professional education training (VET/PET) trainers and other VET/PET professionals at a Swiss institution of higher education, the institution's evaluation unit implemented a design to address questions of measurement invariance across modes. As randomisation on the individual level was not feasible in this setting, randomisation was conducted on the course level. In the data analysis, this issue was addressed statistically. During a given semester, 33 continuing training courses (mostly one-day refresher courses, offered by a different lecturers covering different topics in the field of vocational training) were assigned either to a web survey group ($n_{\text{online}} = 17$) or a paper survey group ($n_{\text{offline}} = 16$). Identical wording and response options for all items in both modes were used to ensure that the layout and visual presentation of the online questionnaire was as close as possible to that used for the paper-and-pencil questionnaire.

For the paper survey group, the administration procedure was the same as the procedure that was routinely implemented previously. The participants completed the paper-and-pencil questionnaire in class at the end of the course. Participants in courses allocated to the web survey group received an email message at the end of the course (8 p.m. on the same day after the course finished) asking them to participate in the evaluation. The invitation email included a personalised link to the online version of the questionnaire. The web survey participants received a second email message 10 days later to remind them to complete the questionnaire if they had not already done so.

To answer the research questions, a CFA was conducted for multiple groups with Mplus7 (Muthén and Muthén 2012). A forward approach with sequentially increasing model constraints, as described by Byrne (2012), was applied. Given that the data were non-normally distributed (Mardia's skewness coefficient = 74.5, $p < 0.01$), maximum likelihood parameter estimates with standard errors (MLR) were used. By adding the 'CLUSTER' command, non-independence was controlled for data at the course level. In addition to testing the models' overall fit, the relative fit of the nested models was tested by calculating the Santorra–Bentler scaled χ^2-difference test. When the test results in a significant difference value, it indicates that adding more equality constraints worsens the model fit (Satorra and Bentler 2010). In addition to χ^2, the comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardised root mean square residual (SRMR) are reported. Decisions about whether the model fit indices were good, acceptable, or not acceptable were based on recommendations made by Bialosiewicz et al. (2013).

4.2 Instrument

The study was carried out for short training courses. To meet the needs of the course lecturers and administrative staff, representatives of these groups were involved in the iterative development of a generic, short questionnaire form. Schnoz-Schmied's (2017) extensive questionnaire, which is empirically well tested and based on an adaptation of Rindermann's (2009) multifactorial model of teaching quality and was used to evaluate the institution's longer training programmes, served as an item pool. The questionnaire construction process resulted in a three-factor structure with 17 items covering the three following aspects:

- Course content, for which there were nine items related to how adequate, informative, helpful, understandable, and goal-oriented the intervention is perceived to be;
- Course lecturers, for which there were five items related to issues like the perception of lecturers' competencies in linking theory with practice and creating a stimulating learning climate; and
- Course administration, for which there were three items covering satisfaction with the course infrastructure and course administration.

Participants were asked to rate the different aspects on a 6-point Likert-type scale ranging from 1 (absolutely disagree) to 6 (absolutely agree). In addition, the following sociodemographic information was included: gender, age, education, and number of years of experience in the professional education field. At the course level, as the courses differed in terms of the topic and lecturer, only course length was included as a control. Table 1 presents Cronbach's α for each scale for both groups separately. The scales were at least acceptable, and mostly excellent, according to Darren and Mallery's (2002) recommendation for interpreting Cronbach's α values.

4.3 Sample

The participants, who attended continuing training courses at an institution of higher education in the German-speaking part of Switzerland, were in their 40s and mainly

Table 1 Cronbach's α for the three scales of the course evaluation questionnaire

Scale	Online	Offline
Course content (9 items)	.90	.89
Course lecturers (5 items)	.90	.86
Course administration (3 items)	.67	.58

male. Generally, they had worked in the field of VET/PET education for more than 10 years, mostly as teachers at VET schools. The participants completed the evaluation questionnaire as a routine procedure for all the training courses provided by the institution.

Data were gathered from 463 participants, of whom 232 completed the online questionnaire and 231 completed the paper-and-pencil questionnaire. Comparison of the response rates of the two groups shows that the average response rate for the online group (84%) was 9% lower than for the offline group (93%). This finding corresponds with most of the former research evidence (see overview on p. 4). Based on the conducted power analysis, we conclude that the sample size of our study is acceptable, although it is important to note that very small effects may not be detected.

A summary of the online and offline samples is given in Table 2. Pearson's chi-square tests, Spearman's rho correlation, and independent-sample *t* tests were used to check whether the two groups were comparable in terms of different sociodemographic variables. An alpha level of .05 for all statistical tests was used. No significant differences were found regarding gender distribution, highest levels of education, mean age, or mean professional experience in years across the two groups. The only significant difference was the overrepresentation of participants who attended courses lasting more than one day in the online sample.

5 Results

5.1 Establishing the baseline models

The questionnaire was developed to assess the quality of course content, course lecturer, and course administration as perceived by the course participants. As a first step, the CFA tested this three-factor structure for both models separately. Since the fit indices for both groups were just acceptable (online: $\chi^2_{[116]} = 254.877$, $p < 0.001$, CFI = 0.92, RMSEA = 0.06, SRMR = 0.07; offline: $\chi^2_{[116]} = 238.63$, $p < 0.001$, CFI = 0.93, RMSEA = 0.07, SRMR = 0.07), in order to identify possible causes of misfit, the Mplus modification indices were evaluated for each model separately.

The results for the online baseline model suggested that allowing a residual covariance between item 2 and item 8 would improve the model fit. As the model fit did improve significantly ($\Delta\chi^2 = 31.701$, $\Delta\text{df} = 1$, $p < 0.01$), and the items relate to the evaluation of one's personal achievements and one's personal learning success, respectively, allowing this residual covariance seemed defensible. Although the model fit was still not higher than 'acceptable' (see Fig. 1), no further model adaptations were performed since no single large source of misfit could be identified and because of the model's parsimony and comparability.

The results for the offline baseline model suggested that residual covariances should be allowed between items 1 and 4 as well as between items 7 and 9 to improve the model fit. It seems that, for the offline group too, some items' content was associated with similar constructs. As the model fit did improve significantly ($\Delta\chi^2 =$

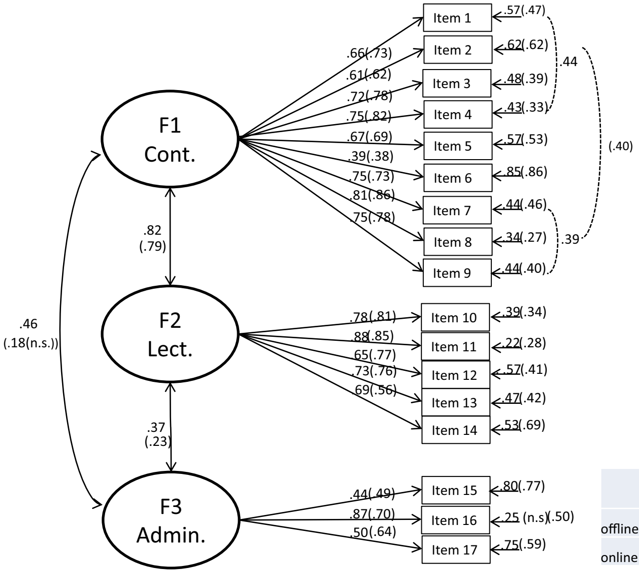
Table 2 Description of the online and offline samples

		Online		Offline		Total		Value	df	Sig.
		N	%	*N*	%	*N*	%			
No. of participants in the evaluation		232	50.1	231	49.9	463	100.0			
No. of courses		17	51.5	16	48.5	33	100.0			
Gender(1):	Women	71	15.4	59	12.8	130	28.1			
	Men	160	34.6	172	27.0	332	71.9	1.541	1	0.214
Course duration (no. of participants and (*no. of courses*))(1):	1 day	129 (*9*)	27.9	177 (*10*)	38.2	306	66.1			
	More than 1 day	103 (*8*)	22.2	54 (*6*)	11.7	157	33.9	22.820	1	< 0.001
Top educational level (no. of participants)(2):	Academic baccalaureate	1	0.2	2	0.5	3	0.7			
	Vocational education	7	1.6	10	2.3	17	3.9			
	Teaching diploma	37	8.6	25	5.8	62	14.4			
	Professional college, etc.	81	18.8	80	18.5	161	37.3			
	Uni. (of applied sciences)	95	22.0	82	19.0	177	41.0			
	PhD	2	0.5	0	0.0	2	0.5			
	Other	7	1.6	3	0.7	10	0.2	− 0.031		0.523
Age in years (mean)(3)		46.5		46.9		46.7		− 0.475	450	0.635
Job-related experience in VET/PET in years (mean)(3)		14.5		15.0		14.7		− 0.513	449	0.608

(1) Pearson’s chi-square test

(2) Spearman’s rho correlation

(3) Independent samples *t* test



	SB-MLRχ2	df	p	RMSEA	CFI	SRMR
offline	191.542	114	0.00	0.055	0.952	0.065
online*	229.987	115	0.00	0.066	0.933	0.064

"Note. Cont. = course content; Lect. = course lecturers; Admin.= course administration"
*=Estimates in brackets are the estimates for the online group

Fig. 1 The two baseline models

24.89, Δdf = 1, $p < 0.01$), these residual covariances were allowed (see Fig. 1). Although no fully identical baseline models for the two groups were identified, according to Byrne (2012), multigroup invariance testing with partially invariant baseline models is a defensible and established strategy.

5.2 Testing dimensional and configural invariance

Dimensional and configural invariance was tested with the configural model. That is, both baseline models were tested in a combined model without adding any equality constraints. The fit indices for the configural model (model 1 in Table 3) were 'acceptable', indicating that the same number of factors and the same item sets for each factor could be assumed across the two modes. In other words, for both the online and offline samples, the three initially conceptualised scales of the course evaluation questionnaire seem to work, and all the items seem to be significantly related to their originally assigned scales.

5.3 Testing metric and scalar invariance

The next step was to test for equal factor loadings across groups. This again resulted in acceptable model fit (model 2 in Table 3). As the model fit did not significantly change according to the Santorra–Bentler scaled χ^2-difference test, the assumption of equivalent loadings across the two modes holds. To test the invariance of the intercepts, these parameters were additionally constrained equal in both groups (model 3 in Table 3). As the model fit was again acceptable and $\Delta\chi^2$ was not significant, it can be assumed that the intercepts too are invariant across modes. These results imply that measurement of the three factors seems to work similarly across modes and that no systematic measurement

Table 3 Results of the MGCFA-models tested

Model	Model description	SB-MLRχ^2	df	CFI (~ > 0.95)	RMSEA (~ < 0.06)	SRMR (~ < 0.08)	Model comparison	$\Delta\chi^{2(*)}$	Δdf	*p*
Configural model										
M1: *configural inv.*	*No constraints*	434.454	229	0.94	0.06	0.07				
Measurement parameters										
M2: *metric inv.*	*Factor loadings invariant*	453.761	243	0.94	0.06	0.10	2 vs. 1	20.930	14	n.s.
M3: *scalar inv.*	Factor loadings invariant; *intercepts invariant*	474.549	260	0.94	0.06	0.11	3 vs. 2	24.840	17	n.s.
Structural parameters										
M4: *structural inv.*	Factor loadings invariant; intercepts invariant; *factor variances and factor covariances invariant*	481.188	266	0.94	0.06	0.16	4 vs. 3	8.229	6	n.s.
Testing for latent mean differences										
M5:	Factor loadings invariant; intercepts invariant; factor variances and factor covariances invariant; *factor means are fixed to 0 for the offline group*	483.869	263	0.94	0.06	0.14		(1)		
Testing for effects of course length										
M6:	Factor loadings invariant; factor variances and factor covariances invariant; factor means are fixed to 0 for the offline group	541.573	291	0.93	0.06	0.13		(2)		

[*] The Satorra-Bentler scaled chi-square difference test was calculated as described on https://stats.idre.ucla.edu/mplus/faq/how-can-i-compute-a-chi-square-test-for-nested-models-with-the-mlr-or-mlm-estimators/

(1) Latent means for the online group: F1 = − 0.239 (n.s.); F2 = − 0.289 (n.s); F3 = 0.113 (n.s.)

(2) Effects of course length on latent means:

on: F1 ON course length = 0.172 (n.s); F2 ON course length = 0.117 (n.s); F3 ON course length = − 0.069 (n.s)

off: F1 ON course length = 0.075 (n.s); F2 ON course length = 0.144 (n.s); F3 ON course length = − 0.017 (n.s) biases are induced by the mode of survey administration, which affect the interpretation of the latent factor means.

5.4 Testing the factor variances and covariances invariance

To test for structural invariance, the factor variances and covariances were constrained equal. As the model fit was again fairly acceptable and model comparison indicated no significant worsening of model fit (model 4 in Table 3), structural invariance was assumed across the modes. In addition, this result suggests that each of the three latent factors explains the same amount of variance in both modes and that the strength of the relationships between the three factors is the same for the online and offline sample.

5.5 Testing for latent mean differences

As a last step, we wanted to know whether the latent factor means differed across the two groups. Since latent factor means are not directly observable, no direct estimate can be made. The standard procedure proposed by Byrne (2012) is to use one group as a reference against which the other group is tested. Therefore, the latent means for the offline group were fixed at 0 and the factor variances were fixed at 1, whereas the parameters for the online group were freely estimated. Again, the model fit was fairly acceptable (model 5 in Table 3). As we found metric and scalar invariance, we conclude that the latent means can be interpreted unambiguously. Therefore, as the latent means in our study do not differ, we conclude that the two groups do not differ in their course evaluations for the three constructs measured.

5.6 Testing for the effects of course length

To ensure that the assumption of factor mean invariance was defensible and not influenced by course length, which was the only control variable that differed significantly for the two samples, course length was added to the model. This additional estimation (model 6 in Table 3) resulted in acceptable model fit, and the loadings of course length on all latent factors were not significant for both modes. This indicates that course length did not influence course evaluations in either the online or offline sample.

6 Interpretation and discussion of the results

The stepwise MGCFA analysis implies invariance across modes when measuring the quality of teaching with the instrument in use, although based on two baseline models that are not completely identical. This indicates that the data collected with paper-and-pencil questionnaires are of equal quality to the data collected by online questionnaires, and therefore, the latent means and correlations between the factors are comparable for both modes.

It seems defensible to assume that the originally conceptualised three-factor structure for the course evaluation questionnaire works for both modes almost equivalently and the items in the online and offline sample load on the same factors. This means that the latent factors cover the same breadth of content in both modes, and therefore, the assumption of dimensional and configural invariance holds. It also seems defensible to assume that no systematic measurement errors resulting from a difference in response style or interpretation of the meaning of one or several items were induced by using an online questionnaire. Therefore, the metric and scalar invariance assumption holds as well. Furthermore, the factors seem to have the same meaning and are equally related to each other across modes, leading us to assume factorial and covariance invariance. As no difference was found for the latent means, we conclude that the two groups do not differ in their evaluations of the three constructs, despite the 9% lower response rate found for the online group. Considering that mode assignment was random at the course level and that the two groups do not differ for various sociodemographic variables, it can be assumed that our findings are not influenced by a selection bias introduced with the new mode.

By analysing online/offline data with a comprehensive method like MGCFA, using randomisation at the course level and employing statistical accounting for the data's non-independency at this level, some of the main methodological critiques of mode research on course evaluation were addressed in this study. We provided further evidence that course evaluation data collected with online questionnaires have no different implications for the interpretation of the measured items or the constructs. It seems that no mode-inherent effects (Hox et al. 2015) are contaminating data quality. Furthermore, the results indicate that comparison of an in-class and an out-of-class setting seems to have no effect on data quality. An alternative argument could be that the confounding context variables, mode (online/offline) and setting (in-class/out-of-class), are cancelling each other out. However, since the research results for both factors tend to point in the same direction (i.e. lower item mean values for the online mode as well as for the out-of-class setting), the two context effects would have to add up and not cancel each other out. On the other hand, a subject not yet addressed in the relevant fields of research (to our knowledge) is the fact that the setting is a combination of the context variable 'lecturer present/not present' and different time points of survey completion (at the end of class/sometime after class). Further studies would be needed to isolate these two effects to show which context variable has which effect and whether this would change the interpretation of the results of this study.

The study results show that MGCFA is a valuable—and, from the perspective of design requirements, feasible—method for testing whether an instrument works the same way in two or more different groups. With the nested step-by-step approach, it can be concretely concluded on which level (quantitative) comparisons are defensible. In addition, because of the possibility of testing the theoretical underlying concept as a whole, this method provides valuable and sophisticated evidence for underlying evaluative processes beyond simply listing differences in the data quality indicators.

Limitations to the generalisability of the findings, especially for the SET discussion, are caused mainly by the sample: Since sample size allows only limited conclusions to be drawn about very small effects. Additionally, the sample does not consist of students in a strict sense, but of adult learners. Furthermore, the instrument used was developed for the institute's special setting and the course format's specific needs. These characteristics could account for the fact that the response rate for both mode groups was relatively high. Perhaps in a classical SET setting, in which students who attend courses throughout a semester and are graded on their work, a selection bias could arise for an online group compared to an in-class paper-and-pencil group. Moreover, the participants in this study, who are in their 40s and have professional lives, probably completed the online questionnaires either at home or at their workplace, whereas students may complete online questionnaires in more diverse contexts with more distractors (Hardré et al. 2012). This may affect data quality in a different way.

As randomisation was performed on the course level instead of the individual level, selection bias cannot be totally excluded. However, the various statistical strategies we conducted to rule out selection bias, like controlling for covariates and non-independence at the course level, do not suggest the presence of such bias.

As the results show, it seems feasible to use the tested instrument for both offline and online course evaluation. Despite the 9% lower response rate for online course evaluations, a complete switch from paper-and-pencil course evaluations to online course evaluations seems defensible, as a response rate of 84% for an online survey is fairly high and there is no indication that systematic biases affected data quality. As discussed above, since students' completion behaviour is probably somewhat different from adult learners and the course format examined in this study is rather specific, the generalisability of these results for SET is limited. Nevertheless, the results confirm the findings of many SET mode studies, which revealed no relevant differences between modes.

Online course evaluation is a well-established practice for assessing teaching quality in higher and further education and will likely continue to be so with the ongoing trends of globalisation and accreditation in higher education institutions. Additionally, an increasing number of students are completing their SET spontaneously on mobile devices. Thus, future research on SET mode effects should address this issue. Because the completion situations for mobile survey-taking are even more diverse and flexible than classic online survey-taking, research in this field should carefully assess the context of survey completion and analyse its effect on data quality with sound research designs and methods of analysis that can provide comprehensive process information.

Funding Open access funding provided by Swiss Federal Institute for Vocational Education and Training (SFIVET).

References

Avery, R. J., Bryant, W. K., Mathios, A., Kang, H., & Bell, D. (2006). Electronic course evaluations. Does an online delivery system influence student evaluations? *Journal of Economic Education, 37*(1), 21–37.

Barkhi, R., & Williams, P. (2010). The impact of electronic media on faculty evaluation. *Assessment & Evaluation in Higher Education, 35*(2), 241–262. https://doi.org/10.1080/02602930902795927.

Bialosiewicz, S., Murphy, K., & Berry, T. (2013). An introduction to measurement invariance testing. Do our measures measure up? The critical role of measurement invariance. Claremont.

Biemer, P. P., De Leeuw, E. D., Eckman, S., Edwards, B., Kreuter, F., Lyberg, L. E., … West, B. T. (2017). Total survey error in practice. Hoboken: Wiley.

Borch, I., Sandvoll, R., & Risør, T. (2020). Discrepancies in purposes of student course evaluations: what does it mean to be "satisfied"? Educational Assessment, Evaluation and Accountability, 83–102. https://doi.org/10.1007/s11092-020-09315-x.

Byrne, B. M. (2012). *Structural equation modeling with Mplus. Basic concepts, applications, and programming*. Hove: Routledge.

Capa-Aydin, Y. (2016). Student evaluation of instruction. Comparison between in-class and online methods. *Assessment & Evaluation in Higher Education, 41*(1), 112–126. https://doi.org/10.1080/02602938.2014.987106.

Champagne, M. V. (2013). Student use of mobile devices in course evaluation. A longitudinal study. *Educational Research and Evaluation, 19*(7), 636–646.

Crews, T. B., & Curtis, D. F. (2011). Online course evaluations. Faculty perspective and strategies for improved response rates. *Assessment & Evaluation in Higher Education, 36*(7), 865–878. https://doi.org/10.1080/02602938.2010.493970.

Darren, G., & Mallery, P. (2002). *SPSS for Windows step by step. A simple guide and reference, 11.0 update* (4th ed.). Boston: Allyn & Bacon.

De Leeuw, E. D., & Hox, J. J. (2011). Internet surveys as part of a mixed-mode design. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the Internet. Advances in applied methods and research strategies* (pp. 45–76). London: Routledge.

Deutskens, E., de Ruyter, K., & Wetzels, M. (2006). An assessment of equivalence between online and mail surveys in service research. *Journal of Service Research, 8*(4), 346–355. https://doi.org/10.1177/1094670506286323.

Dittmann-Domenichini, N., & Halbherr, T. (2015). Disbelief in teaching evaluation. "Mode" does not matter, time and place do. Limassol: Cyprus: Poster presented on the conference of the European Association for Research in Learning and Instruction (EARLI).

Dommeyer, C. J., Baum, P., Hanna, R. W., & Chapman, K. S. (2004). Gathering faculty teaching evaluations by in-class and online surveys: their effects on response rates and evaluations. *Assessment & Evaluation in Higher Education, 29*(5), 611–623. https://doi.org/10.1080/02602930410001689171.

Donzallaz, D. (2010). Qualitätssicherung und Evaluation an Schweizer Hochschulen – methodische Brückenschläge zwischen externen Ansprüchen und internen Realitäten [Quality assurance and evaluation at Swiss universities - methodological bridge building between external demands . LeGes - Gesetzgebung & Evaluation, 1, 33–42.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory factor analysis framework. *Medical Care, 44*(11), 78–94. https://doi.org/10.1097/01.mlr.0000245454.12228.8f.

Hardré, P. L., Crowson, H. M., & Xie, K. (2010). Differential effects of web-based and paper-based administration of questionnaire research instruments in authentic contexts-of-use. *Journal of Educational Computing Research, 42*(1), 103–133.

Hardré, P. L., Crowson, H. M., & Xie, K. (2012). Examining contexts-of-use for web-based and paper-based questionnaires. *Educational and Psychological Measurement, 72*(6), 1015–1038. https://doi.org/10.1177/0013164412451977.

Hessius, J., & Johansson, J. (2015). Smartphone-based evaluations of clinical placements — a useful complement to web-based evaluation tools. *Journal of Educational Evaluation for Health Professions, 12*(55), 1–6.

Hox, J. J., De Leeuw, E. D., & Zijlmans, E. A. O. (2015). Measurement equivalence in mixed mode surveys. *Frontiers in Psychology, 6*, 87. https://doi.org/10.3389/fpsyg.2015.00087.

Kays, K., Gathercoal, K., & Buhrow, W. (2012). Does survey format influence self-disclosure on sensitive question items? *Computers in Human Behavior, 28*, 251–256. https://doi.org/10.1016/j.chb.2011.09.007.

Klausch, T., Hox, J. J., & Schouten, B. (2013). Measurement effects of survey mode on the equivalence of attitudinal rating scale questions. *Sociological Methods & Research, 42*(3), 227–263. https://doi.org/10.1177/0049124113500480.

Klieger, D., Centra, J., Young, J., Holtzman, S., & Kotloff, L. J. (2014). Testing the invariance of interrater reliability between paper-based and online modalities of the SIR II^{TM} student instructional report. Princton. Retrieved from http://search.ets.org/research/contact.html.

Kordts-Freudinger, R., & Geithner, E. (2013). When mode does not matter. Evaluation in class versus out of class. *Educational Research and Evaluation, 19*(7), 605–614.

Layne, B. H., DeCristoforo, J. R., & McGinty, D. (1999). Electronic versus traditional student ratings of instruction. *Research in Higher Education, 40*(2), 221–232. https://doi.org/10.1023/A:1018738731032.

Leung, D. Y. P., & Kember, D. (2005). Comparability of data gathered from evaluation questionnaires on paper and through the internet. *Research in Higher Education, 46*(5), 571–591. https://doi.org/10.1007/s11162-005-3365-3.

Mitchell, O., & Morales, M. (2017). The effect of switching to mandatory online course assessments on response rate and course ratings. *Assessment & Evaluation in Higher Education, 43*(4), 629–639. https://doi.org/10.1080/02602938.2017.1390062.

Morrison, R. (2011). A comparison of online versus traditional student end-of-course critiques in resident courses. *Assessment & Evaluation in Higher Education, 36*(6), 627–641. https://doi.org/10.1080/02602931003632399.

Morrison, K. (2013). Online and paper evaluations of courses. A literature review and a case study. *Educational Research and Evaluation, 19*(7), 585–604.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles: Muthen & Muthen.

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: what can be done? *Assessment & Evaluation in Higher Education, 33*(3), 301–314. https://doi.org/10.1080/02602930701293231.

Rienties, B. (2014). Understanding academics' resistance towards (online) student evaluation. *Assessment & Evaluation in Higher Education, 39*(8), 987–1001. https://doi.org/10.1080/02602938.2014.880777.

Rindermann, H. (2009). Lehrevaluation: Einführung und Überblick zu Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen mit einem Beitrag zur Evaluation computerbasierten Unterrichts. [Student teaching evaluation: introduction and overview of research and practice of course evaluation at universities with a contribution to the evaluation of computer-based teaching (2. Auflage). Landau: Verlag Empirische Pädagogik.

Risquez, A., Vaughan, E., & Murphy, M. (2015). Online student evaluations of teaching. What are we sacrificing for the affordances of technology? *Assessment & Evaluation in Higher Education, 40*(1), 120–134. https://doi.org/10.1080/02602938.2014.890695.

Romppel, M. (2014). Welche Vorzüge haben konfirmatorische Faktorenanalysen im Vergleich zu explorativen Faktorenanalysen? [What are the advantages of confirmatory factor analyses in comparison to exploratory factor analyses?]. *Psychotherapie, Psychosomatik, Medizinische Psychologie, 64*(5), 200–201.

Sass, D. A., & Schmitt, T. A. (2013). Testing measurement and structural invariance. Implications for practice. In T. Teo (Ed.), *Handbook of quantitative methods for educational research* (1st ed., pp. 315–345). Rotterdam: Sense Publishers. https://doi.org/10.1007/978-94-6209-404-8.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika, 75*(2), 243–248.

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance. Review of practice and implications. *Human Resource Management Review, 18*(4), 210–222. https://doi.org/10.1016/j.hrmr.2008.03.003.

Schnoz-Schmied, T. P. (2017). Vers un soutien ciblé au développement de la formation. Education & Fromation [Towards targeted support for training development. The evaluation of modules in higher education], *Revue Education & Formation*, e-307-01, 145–166.

Shih, T.-H., & Fan, X. (2008). Comparing response rates from web and mail surveys. A meta-analysis. *Field Methods, 20*(3), 249–271.

Skedsmo, G. (2020). Assessment and evaluation with clarifying purposes for policy and practice. Educational Assessment, Evaluation and Accountability. https://doi.org/10.1007/s11092-020-09323-x.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching. The state of the art. *Review of Educational Research, 83*(4), 598–642. https://doi.org/10.3102/0034654313496870.

Spooren, P., Vandermoere, F., Vanderstraeten, R., & Pepermans, K. (2017). Exploring high impact scholarship in research on student's evaluation of teaching (SET). *Educational Research Review, 22*(October), 129–141. https://doi.org/10.1016/j.edurev.2017.09.001.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in. *Journal of Consumer Research, 25*(1), 78–107. https://doi.org/10.1086/209528.

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: differences between educational groups in human values measurement. *Quality and Quantity, 43*. https://doi.org/10.1007/s11135-007-9143-x.

Stowell, J. R., Addison, W. E., & Smith, J. L. (2012). Comparison of online and classroom-based student evaluations of instruction. *Assessment & Evaluation in Higher Education, 37*(4), 465–473. https://doi.org/10.1080/02602938.2010.545869.

Treischl, E., & Wolbring, T. (2017). The causal effect of survey mode on students' evaluations of teaching: empirical evidence from three field experiments. Research in Higher Education. https://doi.org/10.1007/s11162-017-9452-4.

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740.

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*(1), 4–70. https://doi.org/10.1177/109442810031002.

Venette, S., Sellnow, D., & McIntyre, K. (2010). Charting new territory: assessing the online frontier of student ratings of instruction. *Assessment & Evaluation in Higher Education, 35*(1), 101–115. https://doi.org/10.1080/02602930802618336.