Check for updates

Evaluation of the influence of wording changes and course type on motivation instrument functioning in chemistry

Regis Komperda, [ID] Kathryn N. Hosbein [ID] and Jack Barbera [ID] *

Increased understanding of the importance of the affective domain in chemistry education research has led to the development and adaptation of instruments to measure chemistry-specific affective traits, including motivation. Many of these instruments are adapted from other fields by using the word 'chemistry' in place of other disciplines or more general 'science' wording. Psychometric evidence is then provided for the functioning of the new adapted instrument. When an instrument is adapted from general language to specific (*e.g.* replacing 'science' with 'chemistry'), an opportunity exists to compare the functioning of the original instrument in the same context as the adapted instrument. This information is important for understanding which types of modifications may have small or large impacts on instrument functioning and in which contexts these modifications may have more or less influence. In this study, data were collected from the online administration of scales from two science motivation instruments in chemistry courses for science majors and for non-science majors. Participants in each course were randomly assigned to view either the science version or chemistry version of the items. Response patterns indicated that students respond differently to different wordings of the items, with generally more favorable response to the science wording of items. Confirmatory factor analysis was used to investigate the internal structure of each instrument, however acceptable data-model fit was not obtained under any administration conditions. Additionally, no discernable pattern could be detected regarding the conditions showing better data-model fit. These results suggest that even seemingly small changes to item wording and administration context can affect instrument functioning, especially if the change in wording affects the construct measured by the instrument. This research further supports the need to provide psychometric evidence of instrument functioning each time an instrument is used and before any comparisons are made of responses to different versions of the instrument.

Introduction

Large-scale empirical evidence continues to support the relationship between the affective domain and general academic success (Richardson *et al.*, 2012) while at the same time other research with a narrower focus has identified relationships between affective characteristics and success in specific academic areas, including science more generally (Glynn *et al.*, 2009; Fortus, 2014) and chemistry specifically (Zusho *et al.*, 2003; Chan and Bauer, 2014; Ferrell *et al.*, 2016). Additionally, it has been shown that many affective characteristics are discipline-specific (Bandura, 1986; Schiefel, 1991). Therefore, to draw meaningful relationships between these characteristics, affective constructs must be operationalized and measured within a

Department of Chemistry, Portland State University, Portland, OR, USA.
E-mail: jbarbera@pdx.edu

given discipline. Much of the research on the role of affect in chemistry has focused on measuring the relationship between motivation and student performance (Black and Deci, 2000; Taasoobshirazi and Glynn, 2009; Chan and Bauer, 2014; González and Paoloni, 2015; Ferrell *et al.*, 2016; Liu *et al.*, 2017). In support of these goals, chemistry-specific motivation instruments have begun to appear in the CER literature (Bauer, 2005; Uzuntiryaki and Aydin, 2009; Ferrell and Barbera, 2015; Salta and Koulougliotis, 2015; Liu *et al.*, 2017).

These chemistry-specific motivation instruments can be broadly classified into two types of development processes. In the first type, the items are written explicitly for the instrument and are often developed based on a specific theoretical framework. This was the process used for development of the College Chemistry Self-Efficacy Scale (CCSS; Uzuntiryaki and Aydin, 2009) using Bandura's (1993) social-cognitive theory. More frequently, chemistry-specific motivation scales or entire instruments are

developed by adapting existing scales from other research areas and replacing terms, such as 'science', 'math', or 'psychology', with 'chemistry'. This process was used to develop the Chemistry Self-Concept Inventory (CSCI; Bauer, 2005), the Academic Motivation Scale – Chemistry (Liu *et al.*, 2017), the Chemistry Motivation Questionnaire (Salta and Koulougliotis, 2015; Hibbard *et al.*, 2016), and the chemistry-specific interest and effort belief scales (Ferrell and Barbera, 2015).

When adapting an existing instrument or individual scales by changing item wordings or using items in a context other than the one in which they were originally developed and tested, it is necessary for the adapted scales to undergo psychometric evaluation to demonstrate that the new chemistry-specific versions are functioning acceptably and evidence should be provided to show that the data obtained from the instrument can be considered valid and reliable (AERA, APA and NCME, 2014; Arjoon *et al.*, 2013) and is therefore providing a true measure of the construct of interest. This type of psychometric evidence has generally been provided when adapting existing motivation instruments for use in chemistry education research (CER), but the data collected usually only come from administration of the chemistry-specific version of the instrument in a specific course type with the goal of demonstrating acceptable instrument functioning in that particular context. What is absent from these studies is a broader understanding of how the functioning of the adapted instrument compares to the functioning of the original instrument with the same student population, or how the adapted instrument functions in student populations that may differ from those in which the original instrument was developed. This information could prove useful to others wanting to adapt existing instruments to fit their specific research or instructional goals and then draw conclusions comparing data from the adapted instrument and the original instrument.

Research goals

One goal of this research is to understand how adapting motivation items from a domain-general context such as science to a domain-specific context such as chemistry (Pajares and Schunk, 2001) affects the functioning of an instrument. Understanding the effects of wording on instrument functioning will provide evidence for the viability of making comparisons of different types of motivation, such as general science motivation and chemistry-specific motivation, measured using the same instrument with changes to item wordings to reflect the type of motivation being measured. A second goal is to understand if changes to the instrument wording affect instrument functioning in the same way in different course types. For example, to see if the chemistry-specific version of a motivation instrument functions equally well in a course designed for science majors as compared to a course designed for non-science majors. Demonstrating equivalent instrument functioning in different course types would provide evidence that comparisons of motivation across course types would be meaningful. For this study, two instruments addressing motivation from different theoretical perspectives were administered as

a science version and a chemistry version in different chemistry courses to see if any patterns of instrument functioning appeared to exist for different wordings and/or course types and to provide psychometric evidence of the validity and reliability of the data obtained from these instruments.

Methods

Motivation instruments

The Science Motivation Questionnaire II (SMQ II; Glynn *et al.*, 2011) is based on Bandura's social-cognitive theory and emphasizes the multi-faceted nature of motivation by measuring distinct yet related aspects of motivation: intrinsic motivation, self-determination, self-efficacy, and extrinsic motivation. The SMQ II has been revised from its previous forms, based on interviews with students and results from earlier factor analyses, to separate extrinsic motivation into two constructs, grade motivation and career motivation, (Glynn and Koballa, 2006; Glynn *et al.*, 2007, 2009). The SMQ II contains a total of 25 items (indicator variables) equally distributed across the five aspects of motivation (latent variables). Since the five latent variables are hypothesized to be distinct yet related constructs, the model of the SMQ II proposed by the developers is that of a correlated five-factor model.

The original development and testing of the various forms of the original SMQ and SMQ II occurred with a population of science and non-science majors enrolled in university-level biology courses (Glynn and Koballa, 2006; Glynn *et al.*, 2007, 2009, 2011). Though the administration context was biology-specific, the SMQ II is worded to address science motivation generally; the biology wording of the instrument was not tested. Based on their results, the developers suggested that it should be possible to create discipline-specific versions (replacing the word 'science' with 'biology', 'chemistry', or 'physics') which would then need to undergo further psychometric analysis (Glynn *et al.*, 2011).

Two instances of the chemistry-specific versions of the SMQ II have been documented in the CER literature: one with US college students (Hibbard *et al.*, 2016) and the other with Greek public secondary school students (Salta and Koulougliotis, 2015). As part of an evaluation of the effectiveness of a flipped general chemistry course sequence, Hibbard *et al.* administered the chemistry-worded instrument (CMQ II) to approximately 60 female students. Means, standard deviations, and Cronbach's alpha values were provided for each of the five motivation scales and an overall alpha value for the entire instrument was reported. Other than an extremely small alpha value for the grade motivation scale (0.13), all alpha values were similar to those provided by the instrument developers (0.81–0.92). No information was provided about the factor structure of the instrument.

Salta and Koulougliotis (2015) translated the CMQ II into Greek and administered it to 330 students aged 14–17. Aspects of items addressing labs were removed as laboratory activities were deemed not applicable to the Greek secondary school context.

As part of their examination of differences in scale means by gender and age, the authors reported means, standard deviations, and Cronbach's alpha values for each of the five motivation scales. Additionally, the correlated five-factor model of the instrument was tested with confirmatory factor analysis (CFA). Data-model fit for both the entire Greek sample and subsets split by age and gender was similar to data-model fit obtained by the original SMQ II developers using the science wording in college biology courses (Glynn *et al.*, 2011). However, these results still do not provide evidence for the five-factor structure of the CMQ II or of the SMQ II when administered to university-level chemistry students in the US, therefore, the SMQ II was chosen as one of the motivation instruments to be investigated in this study.

The other motivation items used in the present study have not been previously published, and were adapted from educational and psychological research pilot studies with undergraduate students using self-determination theory (SDT; Ryan and Deci, 2000; Skinner *et al.*, 2017) to understand the role of science motivation in academic success. Four of the nine SDT items used in this study are indicators of perceived value while the other five are indicators of belonging. Together the nine items are hypothesized to comprise a correlated two-factor model that SDT describes as promoting internalization, which ultimately sustains intrinsic motivation. These items were selected for inclusion in the current study because they represented a theoretically distinct measure of motivation from the SMQ II and provided an opportunity to examine whether both motivation instruments functioned similarly when their wording was changed from science to chemistry-specific.

Participants and data collection

SMQ II items. After obtaining institutional review board (IRB) approval for the study, student participants were recruited by contacting instructors of introductory and general chemistry courses at 6 different colleges and universities primarily in the Western United States and Pacific Northwest. Courses were classified as introductory chemistry courses if the official course description indicated that no prior chemistry coursework either in high school or college was necessary or recommended for enrollment. Courses were classified as general chemistry courses if the official course description indicated that the course was designed for science or engineering majors and had a required or recommended prerequisite of high school chemistry or equivalent. Instructors of the courses were provided with a link to a Qualtrics survey to post to their course website and asked to play a brief video in class in which a research team member described the study and the consent process. No participation incentives were offered to individual students and no identifying student information was collected.

All surveys were open for a non-exam week selected by the instructor between the end of October and the beginning of December 2016. When taking the survey, students were randomly presented with either the science or chemistry wording for all 25 motivation items. The motivation items

were presented in a randomized order followed by demographic questions.

SDT items. The SDT items were collected as part of a larger ongoing project at a single institution in the Pacific Northwest. All responses were from students enrolled in two sections of the first term of the general chemistry sequence taught by two different instructors. On the first page of the survey, students had the option to provide their name so that their course instructor could award a small amount of extra credit for opening the survey, whether or not the student provided any responses on the survey.

The surveys for both course sections were open for the last week of class before final exams. As with the SMQ II items, students saw either the science or the chemistry wording of all SDT items and the items were presented in a randomized order. Demographic information for students responding to the SDT items was obtained from university records as part of the approved IRB for the project.

Analysis

Response patterns. The first stage of the analysis was examination of how students' responses to the SMQ II and SDT motivation items varied based on the aspect of motivation being addressed, the specific wording of the item, and the course in which the student was enrolled. Response distributions for each motivation item were plotted with the R package likert (Version 1.3.6; Bryer and Speerschneider, 2017). Descriptive statistics for each item, wording, and course condition including the mean, standard deviation, median, skewness, and kurtosis were also computed and are provided in Appendix 1. Since the wording of the items was randomized across students within the same course, examination of the response patterns provided an opportunity to focus specifically on how science motivation may differ from chemistry motivation within the same population of students.

Confirmatory factor analysis. While examination of response patterns to individual items on the motivation scales provides some information about specific aspects of general science motivation as compared to chemistry-specific motivation, the most common usage of these types of instruments is to average or sum responses to individual items to create scale scores or overall motivation scores which can then be compared across groups of individuals or administration conditions. For these comparisons to be meaningful, it is necessary to demonstrate that the instrument is functioning well in each context where it was administered. One method of demonstrating acceptable instrument functioning is to provide evidence that the internal structure of the instrument is the same for all groups of responses that will be compared, for example, the science wording of the SMQ II in both general chemistry and introductory chemistry courses.

The first step in examining the internal structure of an instrument is considering potential models for the relationships between the individual items and the underlying constructs the items are designed to measure. For both the SMQ II and SDT items, one potential model is a simple single-factor

model where all items are expected to be associated with a single motivation construct. However, each set of items was developed from a theoretical framework of motivation to have distinct factors. The SMQ II theoretical framework hypothesizes that the 25 items actually measure five distinct, yet related aspects of motivation: intrinsic motivation, self-determination, self-efficacy, grade motivation, and career motivation. Therefore, this proposed model of the SMQ II items can be tested as a correlated five-factor model. For the SDT items, the theoretical framework hypothesizes two distinct aspects of motivation, value and belonging, which can be tested with a correlated two-factor model. Confirmatory factor analysis (CFA) was used to provide evidence for the internal structure of both the SMQ II and SDT items by testing how well the data fit the proposed models for each set of items. All analyses were done with the R package lavaan (Version 0.5-23.1097; Rosseel, 2012). Each model, for each instrument, was tested for each wording within each course, resulting in a total of 12 CFAs.

When conducting CFA, an estimator must be chosen that matches the characteristics of the data. Most CFA studies in the CER literature use the maximum likelihood (ML) estimator (or its robust variant; MLR) which assumes a continuous response scale (Uzuntiryaki and Aydin, 2009; Brandriet *et al.*, 2011, 2013; Xu and Lewis, 2011; Ferrell and Barbera, 2015; Salta and Koulougliotis, 2015; Lastusaari *et al.*, 2016; Villafañe *et al.*, 2016; Bunce *et al.*, 2017; Liu *et al.*, 2017). However, examination of the response patterns and descriptive statistics (Appendix 1) for the SMQ II and SDT items in this study indicated that the data were highly skewed due to a tendency for students to infrequently select response options on the extreme ends of the response scales. Though data collected on a five-point Likert-type scale is often considered to be continuous, there were many SMQ II and SDT items in which students only used four of the response scale options resulting in data that were more categorical in nature than continuous. As a result of the categorical nature of the data and its non-normal distribution, robust diagonally weighted least squares (WLSMV) was chosen as the estimator for the CFAs.

The determination of acceptable CFA data-model fit is typically evaluated by meeting cutoff values for specific fit indices. Fit indices can be categorized into three classes: incremental, parsimonious, and absolute. It is recommended to evaluate fit indices from more than one class (Mueller and Hancock, 2010). Incremental fit indices such as the comparative fit index (CFI) and the Tucker-Lewis index (TLI), range from 0 to 1 where a larger value indicates that the proposed model is a better fit for the data than a null model with no relationships among the individual items. For parsimonious fit indices, such as the root mean square error of approximation (RMSEA), values closer to 0 are better because they indicate a smaller difference between the observed data covariance matrix and the model-implied covariance matrix, while accounting for the complexity of the model. Absolute fit indices function similarly to parsimonious fit indices, though the standardized root mean square residual (SRMR) will always decrease as more parameters are added to the model, regardless of their usefulness.

The cutoff values for fit indices most frequently cited in the CER literature are based on the work of Hu and Bentler (1999) using ML estimation. Hu and Bentler advise cutoffs near 0.95 for the CFI and TLI, near 0.06 for the RMSEA, and near 0.08 for the SRMR. Studies of the WLSMV estimator (Yu, 2002; Beauducel and Herzberg, 2006; Bandalos, 2008) have indicated that the CFI, TLI, and RMSEA tend to indicate a better-fitting model than may actually exist and advocate for more stringent cutoff criteria, especially when the number of response categories is smaller than four. Therefore, for this study, a value of greater than or equal to 0.95 was chosen as an acceptable cutoff for the CFI and TLI and a value at or below 0.05 was chosen as an acceptable cutoff for the RMSEA. The use of the SRMR is not recommended with the WLSMV estimator and as a result, only the CFI, TLI, and RMSEA values will be used to determine acceptable data model fit for this study. Additionally, because the CFI and TLI are both incremental fit indices while the RMESA is a parsimonious fit index, it was deemed necessary for both types of indices to reach the cutoff values to draw a conclusion of acceptable data-model fit. Chi-square values are reported for comparison purposes but not used as indicators of data-model fit (Schermelleh-Engel *et al.*, 2003; Mueller and Hancock, 2008).

Scale reliability

While there are many methods of establishing reliability (AERA, APA and NCME, 2014), for this study internal consistency estimates were utilized for each scale. An overall internal consistency value for all items is only meaningful if the items are shown to be unidimensional, that is associated with a single factor, otherwise it is more appropriate to present internal consistency values for each unidimensional scale, as represented by a single factor within the larger instrument (Cronbach, 1951; Barbera, 2016). The unidimensionality of each motivation scale under each administration condition was tested with a single-factor CFA using the WLSMV estimator and fit index cutoff criteria discussed previously (CFI or TLI $\geq$ 0.95 and RMSEA $\leq$ 0.05).

To determine the suitability of either Cronbach's alpha or omega total as an estimate of scale internal consistency, the single-factor CFA models for each scale were tested under both the less restrictive congeneric model, where the relationships between each item and the factor (loadings) are free to take the value that is the best fit for the data, and also under the more restrictive tau-equivalent model, where all item loadings on the factor are forced to be equivalent (Cho and Kim, 2015; Harshman and Stains, 2017; McNeish, 2017). Alpha was used as an internal consistency estimate for a scale if the tau-equivalent model had acceptable data-model fit. Omega total was used as an internal consistency estimate for a scale if the tau-equivalent model had unacceptable data-model fit but the congeneric model had acceptable data-model fit. If neither model had acceptable data-model fit, a scale was determined not to meet the assumptions necessary to report internal consistency,

therefore no internal consistency estimates were provided. The R package userfriendlyscience (Version 0.6-1; Peters, 2017) was used for alpha and omega calculations with polychoric correlations to account for the ordinal nature of the response scale (Gadermann *et al.*, 2012).

Results & discussion

Data cleaning

Four criteria determined inclusion of responses in the final dataset for each survey type: (1) the student provided informed consent and was age 18 or older, (2) the student selected the correct response on a 'check' item asking for a particular response option to be selected, (3) the student responded to all survey items (25 SMQ II items or 9 SDT items), and (4) for the SMQ II items more than 10 responses were obtained from students in the course, or if the course enrollment was under 40 students, the course response rate was greater than 25%. After cleaning, there were 660 usable responses to the SMQ II items; 287 responses were from general chemistry courses and 373 responses were from introductory chemistry courses. There were 410 usable responses to the SDT items (Table 1). All data cleaning and analysis steps for both datasets were performed in R (Version 3.3.3; R Core Team, 2017).

Participant characteristics

The general chemistry students responding to the SMQ II items were primarily female (54%), white (60%), and majoring in biology (38%) or engineering (26%). The introductory chemistry students responding to the SMQ II items were even more predominantly female (69%) and white (69%), and had an equal proportion of biology majors (31%) and students who described their major as "Other science (not chemistry, biology, or physics)." The general chemistry students responding to the SDT items were primarily female (53%), white (53%), and majoring in biology (29%), engineering (17%), or health studies (14%).

Table 1 Responses to motivation instruments by course type and wording

| Course	Wording	SMQ II ($N = 660$)	SDT ($N = 410$)
General Chemistry	Science	146	208
	Chemistry	141	202
Introductory Chemistry	Science	189	—
	Chemistry	184	—

Note: SDT scales were not administered in introductory courses.

Item response distributions

Examination of the response patterns and descriptive statistics for all motivation items indicated that, in general, more favorable responses to the items, defined as selecting a higher frequency response option for the SMQ II items (usually or always) or an agree response option for the positively worded SDT items and a disagree response to the negatively worded items, were seen for general chemistry students and for students responding to the science wording of the items. Fig. 1 shows the response distributions for general chemistry students on each wording of the first item on the intrinsic motivation scale. The item is written out such that [] indicates where either the word science or chemistry was inserted. The horizontal bars underneath the item wording show the distribution of student responses. The response distributions for each item have been split vertically to allow for comparison of responses to the science wording and the chemistry wording. The percentages listed on the far left of each horizontal bar are the summed percentage of students responding "Never" or "Rarely" and the percentages listed on the far right of each horizontal bar are the summed percentage of students responding "Usually" or "Always." The space representing percentage of students answering "Sometimes" is in the center along with the percentage of responses. For example, Fig. 1 shows that more students responded favorably to the science wording (86%) compared to the chemistry wording (56%) of this item. Similarly, the chemistry wording produced more unfavorable responses (13%) compared to the science wording (4%).

Fig. 2 shows the response distributions for all items on the SMQ II. Response distributions in the left column are from students in the general chemistry courses and response distributions in the right column are from students in the introductory chemistry courses.

Response distributions for general chemistry students on the value and belonging SDT items are shown in Fig. 3. Looking first at the results from general chemistry courses, large differences were apparent in the response patterns for the science wording and chemistry wording of items related to some aspects of motivation, such as intrinsic motivation, career motivation, and belonging. In general, when seeing the science wording of the items, the general chemistry students responded in a way that indicates higher levels of those specific motivation aspects, either by choosing a higher frequency response for the SMQ II items (Fig. 2) or by agreeing more with the positively worded SDT items and disagreeing more with the negatively worded SDT items (Fig. 3). Yet for other aspects of motivation, such as self-determination, self-efficacy, grade motivation, and value, the differences in responses to the two wordings for general chemistry students were less pronounced.
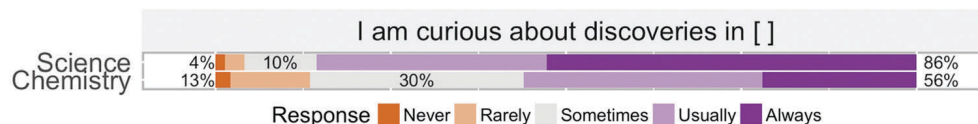
Fig. 1 General chemistry student responses to first intrinsic motivation item on SMQ II by wording.

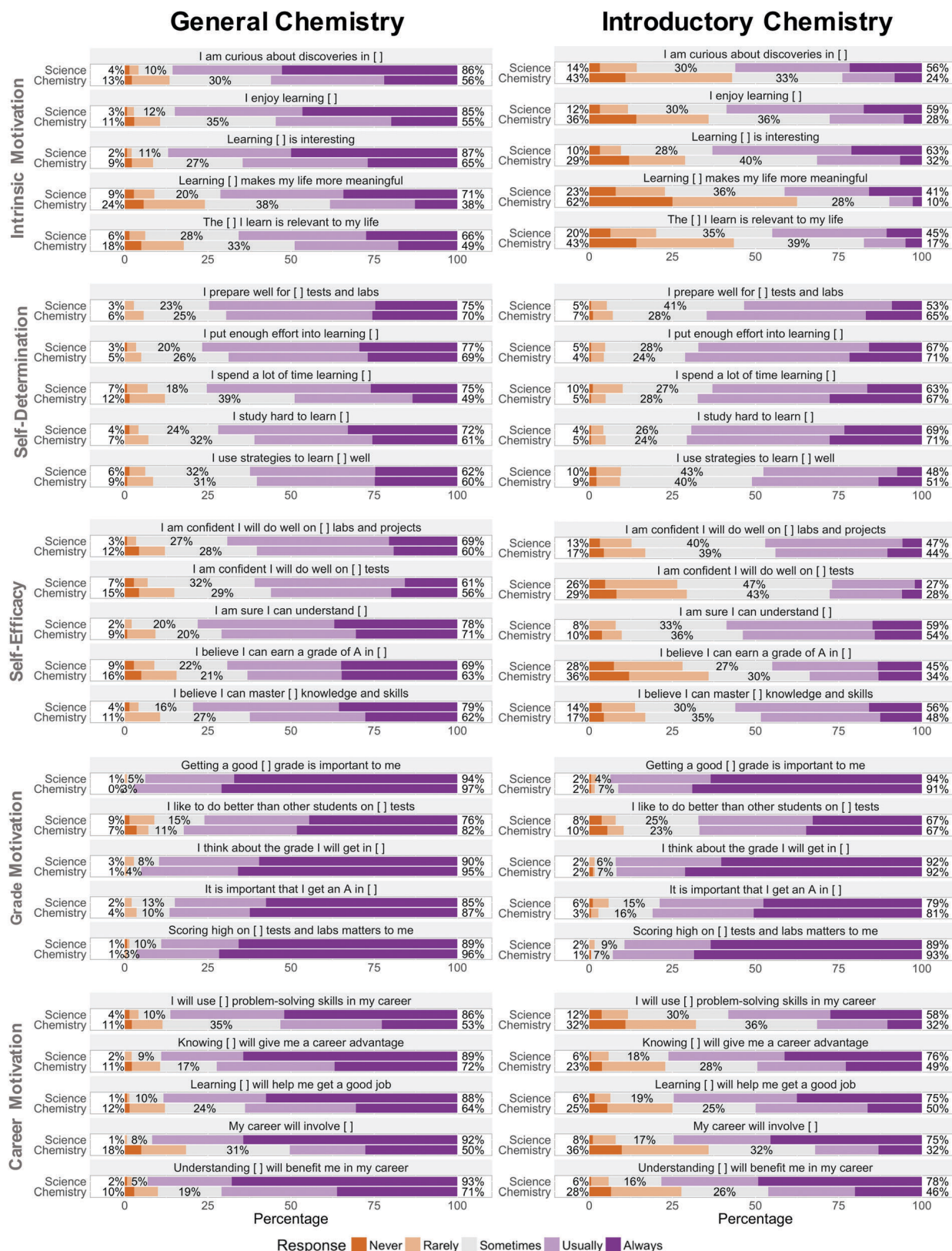Fig. 2 SMQ II item response by wording and by course.

The SMQ II response patterns for the introductory chemistry students were generally similar to the general chemistry students in that large differences based on wording were seen in responses to intrinsic and career motivation items where again higher frequency responses were chosen for the science wording as compared to the chemistry wording. Additionally,
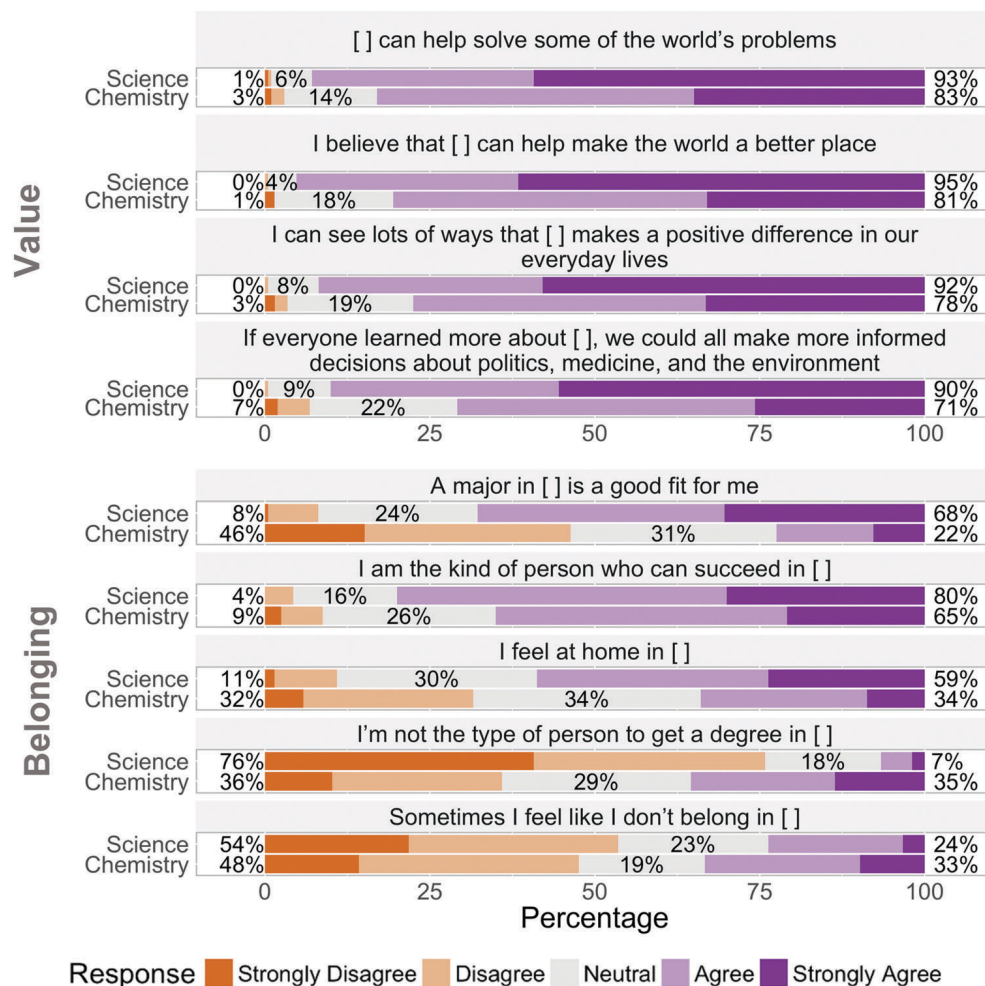
Fig. 3 Response distributions for SDT items from general chemistry students.

both groups of students generally selected higher frequency responses for the chemistry wording of the grade motivation items. However, the introductory chemistry students were overall selecting lower frequency responses to the items than the general chemistry students. Another notable difference between the introductory chemistry students and general chemistry students is that the general chemistry students were more likely to select higher frequency responses to the chemistry worded self-determination items whereas the introductory chemistry students were more likely to select higher frequency responses to the science worded self-determination items.

Internal structure

Following the recommendations of Arjoon *et al.* (2013), two alternative models of the internal structure of both the SMQ II and SDT items were tested using CFA. First, the most parsimonious single-factor model was tested for the SMQ II and SDT items separately for each administration condition. The data-model fit for all single-factor SMQ II and single-factor SDT models (Tables 2 and 3) were unacceptable based on the cutoff criteria used in this study where both the incremental (CFI or TLI) and parsimonious fit indices (RMSEA) must fall within an

acceptable range (CFI or TLI $\geq$ 0.95 and RMSEA $\leq$ 0.05). These results indicate that a single-factor model is not adequate to explain the relationships between either the SMQ II items or the SDT items.

As the single-factor models showed unacceptable fit to the data, multi-factor models corresponding to the theoretical frameworks for the SMQ II and the SDT items were tested. The SMQ II was hypothesized to have an internal structure comprised of five correlated factors representing intrinsic motivation, self-determination, self-efficacy, grade motivation, and career motivation with five items associated with each factor. The SDT items were hypothesized to have an internal structure with two correlated factors representing the value and belonging aspects of SDT with four items and five items associated with each factor, respectively. Again, all multi-factor models were tested for each administration condition.

For both the SMQ II and SDT items, the multi-factor models had better data-model fit than the single-factor models (Tables 4 and 5). However, the data-model fit was unacceptable for all wording conditions and for each course type. In the general chemistry courses, the five-factor model of the SMQ II had worse data-model fit for the science wording responses.

Table 2 Data-model fit for single-factor SMQ II items by course and wording conditions

Course	Wording	χ^2	CFI	TLI	RMSEA
General Chemistry	Science	965	0.80	0.78	0.13
General Chemistry	Chemistry	1577	0.75	0.72	0.18
Introductory Chemistry	Science	1373	0.85	0.84	0.15
Introductory Chemistry	Chemistry	1919	0.77	0.74	0.18

Note. For all models df = 275 and $p < 0.01$.

Table 3 Data-model fit for single-factor SDT items by wording condition

Wording	χ^2	CFI	TLI	RMSEA
Science	361	0.89	0.86	0.24
Chemistry	438	0.84	0.78	0.28

Note. For all models df = 27 and $p < 0.01$.

Table 4 Data-model fit for five-factor SMQ II items by course and wording conditions

Course	Wording	χ^2	CFI	TLI	RMSEA
General Chemistry	Science	483	0.94	0.93	0.08
General Chemistry	Chemistry	468	**0.96**	**0.96**	0.07
Introductory Chemistry	Science	487	**0.97**	**0.97**	0.07
Introductory Chemistry	Chemistry	657	0.94	0.94	0.09

Note. For all models df = 265 and $p < 0.01$; values within acceptable cutoffs are bolded.

Table 5 Data-model fit for two-factor SDT items by wording condition

Wording	χ^2	CFI	TLI	RMSEA
Science	49	**0.99**	**0.99**	0.07
Chemistry	94	**0.97**	**0.96**	0.11

Note. For all models df = 26 and $p < 0.01$; values within acceptable cutoffs are bolded.

For the chemistry wording responses in the general chemistry course the CFI and TLI were above 0.95, but the RMSEA was not below 0.05, indicating unacceptable data-model fit. The opposite pattern of data-model fit was observed for the SMQ II administered in the introductory chemistry courses, though again the models had unacceptable data-model fit in all conditions. In the introductory chemistry courses the science wording responses to the SMQ II met cutoff criteria for the CFI and TLI, but the RMSEA did not meet the cutoff criteria for this study, indicating unacceptable data-model fit.

Though none of the four administration conditions of the SMQ II met the criteria for acceptable data-model fit used in this study with the WLSMV estimator, patterns in the fit indices suggest better fit for the SMQ II items, based on fit indices being closer to the cutoff values used in this study, when the chemistry-worded version was administered to general chemistry students and the science-worded version was administered to introductory chemistry students. The pattern of fit indices for the SDT items administered in the general chemistry courses followed a similar pattern as the SMQ II

administered in introductory chemistry courses where slightly better data-model fit was seen for the science wording responses though, again, under no conditions was data-model fit acceptable. For both wordings of the SDT items the CFI and TLI were above the cutoff values, but the RMSEA never met the cutoff value.

While there are no prior studies with the specific set of SDT items from this research to use as a baseline, the SDT items were chosen for comparison with the SMQ II items since both instruments were developed to address aspects of student motivation. The administration of both instruments to general chemistry students provided an opportunity to look for consistencies in the types of changes occurring as a result of students responding to either the science or chemistry wording. Randomizing the wording seen by students within the same course also provided an opportunity to control for classroom effects that may have been present if the different wordings were presented to different intact courses. Even with this control, the CFA data-model fit for both instruments was unacceptable and did not follow consistent patterns. As with the SMQ II items, while none of the tested models met the data-model cutoff criteria defined for this study, in some conditions the data-model fit indices were closer to acceptable cutoff criteria than in other conditions. The data-model fit for the five-factor SMQ II was better for the chemistry wording in the general chemistry course while the two-factor data-model fit for the SDT items was better for the science wording in the general chemistry course. These results make it difficult to offer any insights to which wording conditions of a motivation instrument are most likely to have acceptable data-model fit in a given course type.

Scale reliability

Given that the multi-factor models for both the SMQ II and SDT items fit better than the single-factor model, there is no evidence that either instrument is unidimensional, which is a fundamental requirement of establishing the internal consistency of scale data. As a result, additional CFAs were conducted to test the unidimensionality of individual motivation scales under both tau-equivalent and congeneric models. The results of these additional CFAs are provided in Appendix 2.

In addition to unidimensionality, Cronbach's alpha has the fundamental assumption that all items are associated with the underlying factor the same degree. This assumption was tested with single-factor tau-equivalent CFA models where all item loadings were constrained to be equal. Of the 24 tau-equivalent models tested, only two showed acceptable data-model fit according to the cutoffs used in this study (CFI or TLI $\geq$ 0.95 and RMSEA $\leq$ 0.05). For these scales, a value of alpha is reported in Table 6.

Omega total is an internal consistency estimate that has the fundamental assumption of unidimensionality, but not of tau-equivalence. Omega allows the items to load to varying degrees when evaluating the single-factor model. This assumption was tested with single-factor congeneric CFA models where the item loadings were not constrained to be equal. Of the 22 scales that did not meet the condition of tau-equivalence, only three scales showed acceptable data-model fit to a single-factor congeneric model. For these scales, a value of omega is reported in Table 6.

Table 6 Scale internal consistency estimates, either omega or *alpha*, by wording and course. A value of NE indicates that no estimate of internal consistency is acceptable for the scale

Scale		Intrinsic motivation		Self-determination		Self-efficacy		Grade motivation		Career motivation		Value		Belonging	
Wording[a]		S	C	S	C	S	C	S	C	S	C	S	C	S	C
Course[b]	GC	NE	NE	NE	NE	NE	NE	NE	0.80	*0.85*	NE	*0.79*	NE	NE	NE
	IC	NE	NE	NE	0.83	NE	NE	0.86	NE	NE	NE	—	—	—	—

[a] S = Science, C = Chemistry. [b] GC = General Chemistry, IC = Introductory Chemistry.

No estimate of internal consistency is provided for scales that did not demonstrate acceptable data-model fit to the less restrictive congeneric model.

The values provided in Table 6 are all above the typically acceptable internal consistency estimates presented in the literature, and similar to values obtained in other studies for the science and chemistry wording of the SMQ II (Glynn *et al.*, 2011; Salta and Koulougliotis, 2015; Hibbard *et al.*, 2016). However, there is no universally agreed upon criterion value for acceptable internal consistency (Arjoon *et al.*, 2013; Taber, 2017). As a result, the internal consistency estimates provided in Table 6 can be used for comparison with prior work, but cannot be used as an absolute indicator of scale or instrument quality.

Conclusions

In this study, the functioning of two motivation instruments was examined under different wording and course conditions. One instrument, the Science Motivation Questionnaire II (SMQ II) was developed based on Bandura's social-cognitive theory and has a documented record of development and testing within the science and chemistry education literature (Glynn and Koballa, 2006; Glynn *et al.*, 2009, 2011; Salta and Koulougliotis, 2015; Hibbard *et al.*, 2016). The second instrument was composed of two scales related to self-determination theory (SDT), and the items had only been used in a limited capacity prior to this study (Skinner *et al.*, 2017). Both instruments were administered as a science-worded version and as a chemistry-worded version. The SMQ II was administered to general chemistry and introductory chemistry courses and the items comprising the SDT scales were administered only in general chemistry courses.

The purpose of this research was to examine the effects that changes in item wording or the type of course in which the instrument was administered have at both the individual item and the overall instrument level. Understanding these effects is necessary to determine the conditions under which the data obtained from an instrument show acceptable evidence for validity and reliability. There are numerous types of evidence that can be provided for the validity and reliability of data, and interested readers are encouraged to consult the *Standards for Educational & Psychological Testing* (AERA, APA and NCME, 2014) and the work of Arjoon *et al.* (2013). In this study, to provide evidence that the motivation constructs being measured, both general science and chemistry-specific, were measured equally well under all conditions validity evidence was examined by testing the internal structure of the SMQ II and SDT items with confirmatory factor analysis (CFA).

The results of the CFAs conducted for this study provide no evidence to support the proposed internal structure of either motivation instrument in any of the conditions tested in this study. This demonstrates that what might appear to be minor changes to the wording of an instrument or the context where it is administered can have an effect on the structural validity of the data generated. In the case of these two motivation instruments, changing the wording from 'science' to 'chemistry' shifts the focus of the items from measuring general motivation to measuring domain-specific motivation. These aspects of motivation represent different constructs (Pajares and Schunk, 2001) and the current forms of the SMQ II and SDT items are not adequately measuring the constructs within the general chemistry and introductory chemistry courses sampled for this research. As a result, any interpretation or further analysis of scale scores or overall instrument scores would be inappropriate and potentially misleading.

The present study highlights that even with an instrument such as the SMQ II, which has undergone extensive development and testing, using the same wording as the original developers or a modification suggested by the developers (changing science to chemistry) can have an influence on the quality of the data obtained when the instrument is administered in different types of chemistry courses. Though the SDT items had a less robust development history, they also demonstrated similar issues with data quality under different administration conditions. It is also important to note that while some individual motivation scales showed acceptable data-model fit during the internal consistency analysis (Appendix 2), the developers did not intend the instruments to be used as individual scales. Instead, the theoretical framework of both instruments hypothesized a relationship among distinct, yet related, aspects of motivation. Since neither motivation instrument showed acceptable evidence for data quality under the conditions of this study, caution should be taken if using the instruments in their current forms under conditions similar to those investigated here unless additional validity or reliability evidence can be provided for the quality of data obtained from administering either wording of the instrument in introductory or general chemistry courses.

Implications for practitioners

Motivation instruments, such as those tested here, can be used to provide insight into students' motivation within a given learning environment. These insights would come from scoring the items that make up each underlying aspect of motivation (*e.g.*, intrinsic motivation, self-determination, *etc.*). To do this, the student response data must first be shown to support the

theoretical framework that groups these items together and justifies that their scores can be determined. The unacceptable data-model fit, obtained across the different administration conditions in this study, preclude us from further interpreting the SMQ II or SDT data or using either to better understand the learning environment. These results highlight the need for examination of the internal structure of an instrument, or other types of validity and reliability, each time an instrument is administered (AERA, APA and NCME, 2014). There is a pervasive misconception that once an instrument has been published in the literature the instrument itself has become validated in some way (Barbera and VandenPlas, 2011) and that data collected from any subsequent administration of the instrument will therefore be equally valid and reliable, even if changes have been made to the wording of the instrument or the context in which it is used. Instead, only if acceptable evidence for instrument functioning (in this case, fit to the *a priori* models) can be provided should instrument scores be reported and used to draw conclusions about the constructs being measured. Based on results of this study, it may be misleading to interpret responses to the SMQ II or SDT items as true indicators of changes in motivation over time, differences in motivation across groups of students or courses, or as evidence for the influence of teaching strategies on students in introductory or general chemistry courses.

Limitations and future work

The intent of this research was to provide information about the quality of data collected from two motivation instruments under different wording and course administration conditions. Data were collected from students in both general and introductory chemistry courses taught by different instructors at different institutions to provide a more generalizable sample from which to draw conclusions. One limitation of this research is that, while the datasets collected from each administration condition were large enough to test the hypothesized internal structure of the instruments with CFA, there were not enough responses collected from different subpopulations to further investigate differences in response patterns and internal structure based on demographic variables such as gender or declared major. Both of these demographic variables had been found to be related to differences in motivation scale scores in previous studies with the SMQ II and CMQ II (Glynn *et al.*, 2011; Salta and Koulougliotis, 2015), though interpretation of these differences is difficult given the reported data-model fit. Future work will focus on collecting larger samples of responses to the motivation items so that these subpopulations of interest can be examined more closely.

A second limitation is the purely quantitative nature of the evidence for data quality presented in this study. Further work with the SMQ II and SDT motivation items will involve interviewing students to better understand how wording, course type, and demographic factors may influence their responses. Additionally, student responses to individual items will be examined in greater detail to investigate the response process validity and to determine if, and how, students' responses change based on wording and how this may impact the intended meaning of an item.

This study represents a first step in understanding how students' motivation may change based on the type of course in which the student is enrolled or when motivation is contextualized as general science or discipline-specific. However, prior to analyzing any differences in motivation based on context, it is necessary to have a functional instrument that can be used to collect valid and reliable data in each setting of interest. The development and testing of such an instrument will be the focus of future work.

Conflicts of interest

There are no conflicts to declare.

Appendices

Appendix 1. Descriptive statistics for SMQ II and SDT items by course and wording conditions

Tables 7 and 8.

Table 7 Descriptive statistics for SMQ II items by course and wording conditions

Item	Course	Wording	Mean	St. dev.	Median	Min	Max	Skew	Kurtosis
The [ ] I learn is relevant to my life	GC	Chemistry	3.44	1.08	3.00	1.00	5.00	−0.32	−0.48
		Science	3.86	0.92	4.00	1.00	5.00	−0.49	−0.13
	IC	Chemistry	2.65	1.03	3.00	1.00	5.00	0.23	−0.32
		Science	3.29	1.04	3.00	1.00	5.00	−0.37	−0.32
Learning [ ] is interesting	GC	Chemistry	3.81	0.98	4.00	1.00	5.00	−0.57	−0.08
		Science	4.34	0.78	4.50	1.00	5.00	−1.19	1.58
	IC	Chemistry	2.97	1.08	3.00	1.00	5.00	−0.21	−0.53
		Science	3.71	0.97	4.00	1.00	5.00	−0.64	0.24
Learning [ ] makes my life more meaningful	GC	Chemistry	3.21	1.07	3.00	1.00	5.00	−0.08	−0.57
		Science	3.94	1.02	4.00	1.00	5.00	−0.85	0.25
	IC	Chemistry	2.25	1.00	2.00	1.00	5.00	0.57	−0.07
		Science	3.26	1.14	3.00	1.00	5.00	−0.20	−0.61
I am curious about discoveries in [ ]	GC	Chemistry	3.62	1.02	4.00	1.00	5.00	−0.34	−0.54
		Science	4.33	0.87	5.00	1.00	5.00	−1.43	2.05
	IC	Chemistry	2.78	1.09	3.00	1.00	5.00	0.31	−0.54
		Science	3.60	1.04	4.00	1.00	5.00	−0.41	−0.43

Table 7 (*continued*)

Item	Course	Wording	Mean	St. dev.	Median	Min	Max	Skew	Kurtosis
I enjoy learning []	GC	Chemistry	3.61	0.98	4.00	1.00	5.00	−0.37	−0.19
		Science	4.28	0.81	4.00	1.00	5.00	-1.09	1.17
	IC	Chemistry	2.83	1.10	3.00	1.00	5.00	−0.06	−0.70
		Science	3.61	0.98	4.00	1.00	5.00	−0.54	0.04
I put enough effort into learning []	GC	Chemistry	3.91	0.85	4.00	2.00	5.00	−0.31	−0.69
		Science	4.02	0.82	4.00	1.00	5.00	−0.64	0.39
	IC	Chemistry	3.88	0.81	4.00	1.00	5.00	−0.46	0.15
		Science	3.78	0.78	4.00	1.00	5.00	−0.39	0.21
I use strategies to learn [] well	GC	Chemistry	3.76	0.94	4.00	1.00	5.00	−0.28	−0.63
		Science	3.79	0.92	4.00	1.00	5.00	−0.39	−0.20
	IC	Chemistry	3.53	0.89	4.00	1.00	5.00	−0.29	0.13
		Science	3.43	0.82	3.00	1.00	5.00	−0.36	0.45
I spend a lot of time learning []	GC	Chemistry	3.49	0.91	3.00	1.00	5.00	−0.14	−0.32
		Science	3.94	0.86	4.00	1.00	5.00	−0.71	0.31
	IC	Chemistry	3.90	0.88	4.00	1.00	5.00	−0.38	−0.42
		Science	3.68	0.89	4.00	1.00	5.00	−0.47	−0.09
I prepare well for [] tests and labs	GC	Chemistry	3.89	0.85	4.00	2.00	5.00	−0.35	−0.58
		Science	3.97	0.76	4.00	2.00	5.00	−0.31	−0.41
	IC	Chemistry	3.73	0.85	4.00	1.00	5.00	−0.48	0.20
		Science	3.57	0.75	4.00	1.00	5.00	−0.11	0.09
I study hard to learn []	GC	Chemistry	3.79	0.91	4.00	2.00	5.00	−0.16	−0.92
		Science	3.99	0.90	4.00	1.00	5.00	−0.67	0.23
	IC	Chemistry	3.93	0.86	4.00	1.00	5.00	−0.48	−0.20
		Science	3.88	0.83	4.00	1.00	5.00	−0.39	−0.10
I am confident I will do well on [] tests	GC	Chemistry	3.57	1.06	4.00	1.00	5.00	−0.48	−0.31
		Science	3.67	0.89	4.00	1.00	5.00	−0.61	0.67
	IC	Chemistry	2.96	1.00	3.00	1.00	5.00	−0.06	−0.32
		Science	2.98	0.86	3.00	1.00	5.00	−0.21	−0.20
I am confident I will do well on [] labs and projects	GC	Chemistry	3.63	1.02	4.00	1.00	5.00	−0.64	0.10
		Science	3.86	0.80	4.00	1.00	5.00	−0.39	0.18
	IC	Chemistry	3.33	0.97	3.00	1.00	5.00	−0.31	−0.14
		Science	3.37	0.86	3.00	1.00	5.00	−0.53	0.38
I believe I can master [] knowledge and skills	GC	Chemistry	3.79	0.97	4.00	2.00	5.00	−0.29	−0.95
		Science	4.10	0.87	4.00	1.00	5.00	−0.94	1.03
	IC	Chemistry	3.40	1.00	3.00	1.00	5.00	−0.36	−0.24
		Science	3.54	1.00	4.00	1.00	5.00	−0.51	−0.08
I believe I can earn a grade of A in []	GC	Chemistry	3.77	1.18	4.00	1.00	5.00	−0.67	−0.49
		Science	3.92	1.03	4.00	1.00	5.00	−0.79	0.08
	IC	Chemistry	2.99	1.21	3.00	1.00	5.00	0.06	−0.89
		Science	3.23	1.14	3.00	1.00	5.00	−0.22	−0.82
I am sure I can understand []	GC	Chemistry	3.91	0.95	4.00	1.00	5.00	−0.62	−0.30
		Science	4.13	0.80	4.00	2.00	5.00	−0.48	−0.64
	IC	Chemistry	3.54	0.94	4.00	1.00	5.00	−0.52	0.35
		Science	3.66	0.83	4.00	2.00	5.00	−0.14	−0.55
I like to do better than other students on [] tests	GC	Chemistry	4.20	1.01	4.00	1.00	5.00	−1.44	1.79
		Science	4.10	1.01	4.00	1.00	5.00	−0.97	0.15
	IC	Chemistry	3.86	1.12	4.00	1.00	5.00	−0.87	0.17
		Science	3.88	1.04	4.00	1.00	5.00	−0.80	0.27
Getting a good [] grade is important to me	GC	Chemistry	4.68	0.53	5.00	3.00	5.00	−1.34	0.81
		Science	4.60	0.63	5.00	2.00	5.00	−1.47	1.72
	IC	Chemistry	4.58	0.72	5.00	1.00	5.00	−1.90	3.96
		Science	4.54	0.70	5.00	1.00	5.00	−1.85	4.36
It is important that I get an A in []	GC	Chemistry	4.45	0.81	5.00	2.00	5.00	−1.39	1.07
		Science	4.40	0.79	5.00	2.00	5.00	−1.09	0.22
	IC	Chemistry	4.28	0.85	5.00	1.00	5.00	−0.99	0.39
		Science	4.20	0.94	4.00	1.00	5.00	−1.05	0.50

Table 7 (*continued*)

Item	Course	Wording	Mean	St. dev.	Median	Min	Max	Skew	Kurtosis
I think about the grade I will get in []	GC	Chemistry	4.60	0.61	5.00	2.00	5.00	−1.44	1.84
		Science	4.47	0.75	5.00	2.00	5.00	−1.37	1.41
	IC	Chemistry	4.60	0.73	5.00	1.00	5.00	−2.23	5.91
		Science	4.51	0.69	5.00	2.00	5.00	−1.34	1.49
Scoring high on [] tests and labs matters to me	GC	Chemistry	4.67	0.57	5.00	2.00	5.00	−1.75	3.24
		Science	4.53	0.75	5.00	1.00	5.00	−1.68	2.92
	IC	Chemistry	4.60	0.66	5.00	1.00	5.00	−1.85	4.30
		Science	4.51	0.73	5.00	2.00	5.00	−1.37	1.16
Learning [] will help me get a good job	GC	Chemistry	3.81	1.03	4.00	1.00	5.00	−0.50	−0.61
		Science	4.44	0.76	5.00	1.00	5.00	−1.38	2.10
	IC	Chemistry	3.36	1.13	3.50	1.00	5.00	−0.28	−0.81
		Science	4.04	0.95	4.00	1.00	5.00	−0.86	0.35
Knowing [] will give me a career advantage	GC	Chemistry	3.96	1.04	4.00	1.00	5.00	−0.84	0.00
		Science	4.51	0.74	5.00	2.00	5.00	−1.44	1.35
	IC	Chemistry	3.46	1.15	3.00	1.00	5.00	-0.20	−0.95
		Science	4.11	0.92	4.00	1.00	5.00	−0.79	−0.08
Understanding [] will benefit me in my career	GC	Chemistry	3.94	1.05	4.00	1.00	5.00	−0.85	0.11
		Science	4.58	0.71	5.00	1.00	5.00	−2.05	5.12
	IC	Chemistry	3.32	1.20	3.00	1.00	5.00	−0.16	−0.99
		Science	4.21	0.93	4.00	1.00	5.00	−0.98	0.12
My career will involve []	GC	Chemistry	3.55	1.17	4.00	1.00	5.00	−0.31	−0.83
		Science	4.55	0.66	5.00	2.00	5.00	−1.32	1.01
	IC	Chemistry	2.99	1.17	3.00	1.00	5.00	0.13	−0.82
		Science	4.11	1.00	4.00	1.00	5.00	−0.90	−0.05
I will use [] problem−solving skills in my career	GC	Chemistry	3.62	1.00	4.00	1.00	5.00	−0.26	−0.51
		Science	4.33	0.86	5.00	1.00	5.00	−1.44	2.20
	IC	Chemistry	2.99	1.13	3.00	1.00	5.00	0.00	−0.67
		Science	3.70	1.07	4.00	1.00	5.00	−0.51	−0.33

Note: GC = General chemistry; IC = Introductory chemistry.

Table 8 Descriptive statistics for SDT items by wording conditions

Item	Wording	Mean	St. dev.	Median	Min	Max	Skew	Kurtosis
I believe that [] can help make the world a better place	Chemistry	4.12	0.77	4.00	1.00	5.00	−0.72	1.05
	Science	4.56	0.60	5.00	2.00	5.00	−1.16	0.95
I can see lots of ways that [] makes a positive difference in our everyday lives	Chemistry	4.06	0.86	4.00	1.00	5.00	−0.87	0.99
	Science	4.50	0.66	5.00	2.00	5.00	−1.04	0.33
[] can help solve some of the world's problems	Chemistry	4.15	0.77	4.00	1.00	5.00	−0.78	0.77
	Science	4.50	0.68	5.00	1.00	5.00	−1.48	3.01
If everyone learned more about [], we could all make more informed decisions about politics, medicine, and the environment	Chemistry	3.90	0.91	4.00	1.00	5.00	−0.78	0.63
	Science	4.45	0.69	5.00	2.00	5.00	−0.93	−0.02
A major in [] is a good fit for me	Chemistry	2.70	1.13	3.00	1.00	5.00	0.31	−0.60
	Science	3.89	0.94	4.00	1.00	5.00	−0.47	−0.57
I am the kind of person who can succeed in []	Chemistry	3.77	0.93	4.00	1.00	5.00	−0.67	0.42
	Science	4.06	0.80	4.00	2.00	5.00	−0.62	0.01
I feel at home in []	Chemistry	3.07	1.04	3.00	1.00	5.00	0.05	−0.64
	Science	3.70	0.98	4.00	1.00	5.00	−0.34	−0.52
Sometimes I feel like I don't belong in []	Chemistry	2.82	1.22	3.00	1.00	5.00	0.21	−1.06
	Science	2.52	1.14	2.00	1.00	5.00	0.27	−0.97
I'm not the type of person to get a degree in []	Chemistry	3.01	1.19	3.00	1.00	5.00	0.08	−0.91
	Science	1.93	0.97	2.00	1.00	5.00	0.96	0.50

Appendix 2. Congeneric and tau-equivalent data-model fit for SMQ II and SDT scales

Tables 9–12.

Table 9 Congeneric data-model fit for each SMQ II scale by course and wording conditions

Scale	Course	Wording	χ^2	p	CFI	TLI	RMSEA
Intrinsic motivation	GC	Science	21.7	<0.01	**0.98**	**0.97**	0.15
		Chemistry	61.4	<0.01	**0.96**	0.92	0.28
	IC	Science	10.2	0.07	**1.00**	**1.00**	**0.08**
		Chemistry	49.3	<0.01	**0.98**	**0.96**	0.22
Self-determination	GC	Science	16.9	<0.01	**0.98**	0.95	0.13
		Chemistry	8.7	0.12	**1.00**	**1.00**	**0.07**
	IC	Science	17.6	<0.01	**0.98**	**0.97**	0.12
		Chemistry	7.2	0.20	**1.00**	**1.00**	**0.05**
Self-efficacy	GC	Science	13.4	0.02	**0.99**	**0.98**	0.11
		Chemistry	15.3	<0.01	**0.99**	**0.99**	0.12
	IC	Science	10.0	0.07	**1.00**	**0.99**	**0.07**
		Chemistry	18.5	<0.01	**0.99**	**0.98**	0.12
Grade motivation	GC	Science	10.1	0.07	**0.99**	**0.99**	**0.08**
		Chemistry	2.5	0.78	**1.00**	**1.01**	**0.00**
	IC	Science	6.4	0.27	**1.00**	**1.00**	**0.04**
		Chemistry	9.5	0.09	**0.99**	**0.99**	**0.07**
Career motivation	GC	Science	7.3	0.20	**1.00**	**1.00**	**0.06**
		Chemistry	12.2	0.03	**1.00**	**0.99**	0.10
	IC	Science	21.0	<0.01	**1.00**	**0.99**	0.13
		Chemistry	52.6	<0.01	**0.99**	**0.98**	0.23

Note: for all models df = 5 and WLSMV estimator used; values within acceptable cutoffs are bolded.

Table 10 Tau-equivalent data-model fit for each SMQ II scale by course and wording conditions

Scale	Course	Wording	χ^2	p	CFI	TLI	RMSEA
Intrinsic motivation	GC	Science	29.2	<0.01	**0.98**	**0.98**	0.12
		Chemistry	94.5	<0.01	0.94	0.93	0.26
	IC	Science	88.3	<0.01	**0.96**	**0.96**	0.22
		Chemistry	83.6	<0.01	**0.97**	**0.96**	0.21
Self-determination	GC	Science	29.1	<0.01	**0.96**	**0.96**	0.12
		Chemistry	38.8	<0.01	**0.98**	**0.98**	0.15
	IC	Science	36.0	<0.01	**0.96**	**0.96**	0.13
		Chemistry	61.0	<0.01	**0.96**	0.95	0.18
Self-efficacy	GC	Science	22.3	0.01	**0.98**	**0.98**	0.10
		Chemistry	15.6	0.08	**1.00**	**1.00**	**0.07**
	IC	Science	57.6	<0.01	**0.97**	**0.97**	0.17
		Chemistry	17.0	0.05	**0.99**	**0.99**	**0.07**
Grade motivation	GC	Science	54.6	<0.01	0.95	0.94	0.19
		Chemistry	61.3	<0.01	0.94	0.93	0.20
	IC	Science	44.7	<0.01	**0.96**	**0.96**	0.15
		Chemistry	35.2	<0.01	**0.96**	0.95	0.13
Career motivation	GC	Science	11.1	0.27	**1.00**	**1.00**	**0.04**
		Chemistry	38.5	<0.01	**0.99**	**0.99**	0.15
	IC	Science	44.6	<0.01	**0.99**	**0.99**	0.15
		Chemistry	85.6	<0.01	**0.98**	**0.98**	0.22

Note: for all models df = 9 and WLSMV estimator used; values within acceptable cutoffs are bolded.

Table 11 Congeneric data-model fit for each SDT scale by wording

Scale	Wording	χ^2	df	p	CFI	TLI	RMSEA
Value	Science	5.6	2	0.06	**1.00**	**1.00**	0.09
	Chemistry	9.6	2	<0.01	**1.00**	**0.99**	0.14
Belonging	Science	11.3	5	0.05	**1.00**	**0.99**	0.08
	Chemistry	49.2	5	<0.01	**0.96**	0.93	0.21

Note: WLSMV estimator used; values within acceptable cutoffs are bolded.

Table 12 Tau-equivalent data-model fit for each SDT scale by wording

Scale	Wording	χ^2	df	p	CFI	TLI	RMSEA
Value	Science	7.4	5	0.19	**1.00**	**1.00**	**0.05**
	Chemistry	70.6	5	<0.01	**0.96**	0.95	0.26
Belonging	Science	51.1	9	<0.01	**0.97**	**0.97**	0.15
	Chemistry	50.9	9	<0.01	**0.97**	**0.96**	0.15

Note: WLSMV estimator used; values within acceptable cutoffs are bolded.

Acknowledgements

References

American Educational Research Association (AERA), American Psychological Association (APA) and National Council on Measurement in Education (NCME), (2014), *Standards for educational & psychological testing*, Washington, DC: American Educational Research Association.

Arjoon J. A., Xu X. and Lewis J. E., (2013), Understanding the state of the art for measurement in chemistry education research: examining the psychometric evidence, *J. Chem. Educ.*, **90**(5), 536–545.

Bandalos D. L., (2008), Is parceling really necessary? A comparison of results from item parceling and categorical variable methodology, *Struct. Equ. Modeling*, **15**(2), 211–240.

Bandura A., (1986), *Social foundations of thought and action: a social cognitive theory*, Englewood Cliffs, NJ: Prentice Hall.

Bandura A., (1993), Perceived self-efficacy in cognitive development and functioning, *Educ. Psychol.*, **28**(2), 117–148.

Barbera J., (2016), Is Cronbach's alpha the correct measure of internal consistency reliability for your data? Biennial Conference on Chemical Education, Greely, CO.

Barbera J. and VandenPlas J. R., (2011), All assessment materials are not created equal: the myths about instrument development, validity, and reliability, in Bunce D. M. (ed.), *Investigating classroom myths through research on teaching and learning*, Washington, DC: American Chemical Society, pp. 177–193.

Bauer C. F., (2005), Beyond ''student attitudes'': chemistry self-concept inventory for assessment of the affective component of student learning, *J. Chem. Educ.*, **82**(12), 1864–1870.

Beauducel A. and Herzberg P. Y., (2006), On the performance of maximum likelihood *versus* means and variance adjusted weighted least squares estimation in CFA, *Struct. Equ. Modeling*, **13**(2), 186–203.

Black A. E. and Deci E. L., (2000), The effects of instructors' autonomy support and students' autonomous motivation on learning organic chemistry: a self-determination theory perspective, *Sci. Educ.*, **84**(6), 740–756.

Brandriet A. R., Xu X., Bretz S. L. and Lewis J. E., (2011), Diagnosing changes in attitude in first-year college chemistry students with a shortened version of Bauer's semantic differential, *Chem. Educ. Res. Pract.*, **12**(2), 271.

Brandriet A. R., Ward R. M. and Bretz S. L., (2013), Modeling meaningful learning in chemistry using structural equation modeling, *Chem. Educ. Res. Pract.*, **14**(4), 421–430.

Bryer J. and Speerschneider K., (2017), *likert: Functions to analyze and visualize likert type items*, [Computer software].

Bunce D. M., Komperda R., Schroeder M. J., Dillner D. K., Lin S., Teichert M. A. and Hartman J. R., (2017), Differential Use of Study Approaches by Students of Different Achievement Levels, *J. Chem. Educ.*, **94**(10), 1415–1424.

Chan J. Y. K. and Bauer C. F., (2014), Identifying at-risk students in general chemistry *via* cluster analysis of affective characteristics, *J. Chem. Educ.*, **91**(9), 1417–1425.

Cho E. and Kim S., (2015), Cronbach's coefficient alpha: well known but poorly understood, *Organ. Res. Methods*, **18**(2), 207–230.

Cronbach L. J., (1951), Coefficient alpha and the internal structure of tests, *Psychometrika*, **16**(3), 297–334.

Ferrell B. and Barbera J., (2015), Analysis of students' self-efficacy, interest, and effort beliefs in general chemistry, *Chem. Educ. Res. Pract.*, **318**(16), 318–337.

Ferrell B., Phillips M. M. and Barbera J., (2016), Connecting achievement motivation to performance in general chemistry, *Chem. Educ. Res. Pract.*, **17**(4), 1054–1066.

Fortus D., (2014), Attending to affect, *J. Res. Sci. Teach.*, **51**(7), 821–835.

Gadermann A. M., Guhn M. and Zumbo B. D., (2012), Estimating ordinal reliability for Likert-type and ordinal item response data: a conceptual, empirical, and practical guide, *Pract. Assessment, Res. Eval.*, **17**(3), 1–13.

Glynn S. M. and Koballa T. R., (2006), Motivation to learn in college science, in Mintzes J. J. and Leonard W. H. (ed.), *Handbook of college science teaching*, Arlington, VA: National Science Teachers Association, pp. 25–32.

Glynn S. M., Taasoobshirazi G. and Brickman P., (2007), Nonscience majors learning science: a theoretical model of motivation, *J. Res. Sci. Teach.*, **44**(8), 1088–1107.

Glynn S. M., Taasoobshirazi G. and Brickman P., (2009), Science motivation questionnaire: construct validation with nonscience majors, *J. Res. Sci. Teach.*, **46**(2), 127–146.

Glynn S. M., Brickman P., Armstrong N. and Taasoobshirazi G., (2011), Science motivation questionnaire II: validation with science majors and nonscience majors, *J. Res. Sci. Teach.*, **48**(10), 1159–1176.

González A. and Paoloni P.-V., (2015), Perceived autonomy-support, expectancy, value, metacognitive strategies and performance in chemistry: a structural equation model in undergraduates, *Chem. Educ. Res. Pract.*, **16**(3), 640–653.

Harshman J. and Stains M., (2017), A review and evaluation of the internal structure and consistency of the Approaches to Teaching Inventory, *Int. J. Sci. Educ.*, **39**(7), 1–19.

Hibbard L., Sung S. and Wells B., (2016), Examining the effectiveness of a semi-self-paced flipped learning format in a college general chemistry sequence, *J. Chem. Educ.*, **93**(1), 24–30.

Hu L. and Bentler P. M., (1999), Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria *versus* new alternatives, *Struct. Equ. Modeling*, **6**(1), 1–55.

Lastusaari M., Laakkonen E. and Murtonen M., (2016), ChemApproach: validation of a questionnaire to assess the learning approaches of chemistry students, *Chem. Educ. Res. Pract.*, **17**(4), 723–730.

Liu Y., Ferrell B., Barbera J. and Lewis J. E., (2017), Development and evaluation of a chemistry-specific version of the academic motivation scale (AMS-Chemistry). *Chem. Educ. Res. Pract.*, **18**(1), 191–213.

McNeish D., (2017), Thanks coefficient alpha, we'll take it from here, *Psychol. Methods*, DOI: 10.1037/met0000144.

Mueller R. O. and Hancock G. R., (2008), Best practices in structural equation modeling, in Osborne J. W. (ed.), *Best practices in quantitative methods*, Thousand Oaks, CA: Sage Publications, Inc., pp. 488–508.

Mueller R. O. and Hancock G. R., (2010), Structural equation modeling, in Hancock G. R. and Mueller R. O. (ed.), *The reviewer's guide to quantitative methods in the social sciences*, New York: Routledge, pp. 371–383.

Pajares F. and Schunk D. H., (2001), Self-beliefs and school success: self-efficacy, self-concept, and school achievement, in Riding R. and Rayner S. (ed.), *Perception*, London: Ablex Publishing, pp. 239–266.

Peters G., (2017), *userfriendlyscience: quantitative analysis made accessible*, [Computer software].

R Core Team, (2017), *R: a language and environment for statistical computing*, [Computer software].

Richardson M., Abraham C. and Bond R., (2012), Psychological correlates of university students' academic performance: a systematic review and meta-analysis, *Psychol. Bull.*, **138**(2), 353–387.

Rosseel Y., (2012), lavaan: an R package for structural equation modeling, *J. Stat. Softw.*, **48**(2), 1–36.

Ryan R. and Deci E. L., (2000), Self-determination theory and the facilitation of intrinsic motivation, *Am. Psychol.*, **55**(1), 68–78.

Salta K. and Koulougliotis D., (2015), Assessing motivation to learn chemistry: adaptation and validation of Science Motivation Questionnaire II with Greek secondary school students, *Chem. Educ. Res. Pract.*, **16**(2), 237–250.

Schermelleh-Engel K., Moosbrugger H. and Müller H., (2003), Evaluating the fit of structural equation models: tests of

significance and descriptive goodness-of-fit measures, *Methods Psychol. Res. Online*, **8**(2), 23–74.

Schiefel U., (1991), Interest, Learning, and Motivation, *Educ. Psychol.*, **26**(3 & 4), 299–323.

Skinner E., Saxton E., Currie C. and Shusterman G., (2017), A motivational account of the undergraduate experience in science: brief measures of students' self-system appraisals, engagement in coursework, and identity as a scientist, *Int. J. Sci. Educ.*, DOI: 10.1080/09500693.2017.1387946.

Taasoobshirazi G. and Glynn S. M., (2009), College students solving chemistry problems: a theoretical model of expertise, *J. Res. Sci. Teach.*, **46**(10), 1070–1089.

Taber K. S., (2017), The use of Cronbach's alpha when developing and reporting research instruments in science education, *Res. Sci. Educ.*, DOI: 10.1007/s11165-016-9602-2.

Uzuntiryaki E. and Aydin Y. Ç., (2009), Development and validation of chemistry self-efficacy scale for college students, *Res. Sci. Educ.*, **39**(4), 539–551.

Villafañe S. M., Xu X. and Raker J. R., (2016), Self-efficacy and academic performance in first-semester organic chemistry: testing a model of reciprocal causation, *Chem. Educ. Res. Pract.*, **17**(4), 973–984.

Xu X. and Lewis J. E., (2011), Refinement of a chemistry attitude measure for college students, *J. Chem. Educ.*, **88**(5), 561–568.

Yu C.-Y., (2002), *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continous outcomes*, (Doctoral dissertation), Los Angeles: University of California.

Zusho A., Pintrich P. R. and Coppola B., (2003), Skill and will: the role of motivation and cognition in the learning of college chemistry, *Int. J. Sci. Educ.*, **25**(9), 1081–1094.