International Journal of Teaching and Learning in Higher Education 2022, Volume 33, Number 3, 407-428
http://www.isetl.org/ijtlhe/ ISSN 1812-9129

Development and Validation of a Formative Evaluation Instrument for College Teaching

Ren Liu, Xiufeng Liu, and Lara Hutson
University of Buffalo

This study aims to provide preliminary validation for a newly designed instrument to evaluate teaching effectiveness through student classroom engagement and learning gains. The instrument is titled the Middle Semester Classroom Survey of Student Engagement and Learning (MS-CSSEL); it consists of 31items to measure student classroom engagement in three dimensions and 19 items to measure student learning gains in three dimensions. To validate the instrument, 634 undergraduate students in a four-year research university participated in this study. The multidimensional Rasch model was used to conduct the analysis. The findings indicated that (a) items displayed a good fit to the Rasch model; (b) dimensions were distinct from each other; and (c) the items displayed high reliability. This instrument measures teaching effectiveness in a new perspective and provides college teachers with a new tool to gauge student engagement and learning gains for conducting evidence-based teaching improvement.

Student ratings of instruction (SRI) is among the most predominant approaches adopted by higher education institutions to measure overall teaching performance or effectiveness (Berk, 2005; Chen & Hoshower, 2003; Yao & Grady, 2005; Zabaleta, 2007). Typically, SRI are administrated near the end of each semester, and students are asked to rate the characteristics of teachers and courses, such as course organization, teachers' enthusiasm, and clarity of explanation (Uttl et al., 2017).

Although initially SRI was intended to provide formative feedback to improve teaching quality, since the 1970s, it has been commonly used for making high-stakes decisions for faculty members (Berk,2005; Clayson, 2013; Clayson & Haley, 2011; Galbraith et al., 2012; Nasser & Fresko, 2002; Stowell et al., 2012). However, the original formative purpose of SRI has not been fully achieved. There is little evidence to support the usefulness of SRI to improve and shape the quality of teaching. Faculty members rarely make changes in their teaching styles or course content based on students' ratings, especially not senior faculty (Yao & Grady, 2005). Many instructors are not well trained formally in pedagogy and do not have the necessary skills to interpret students' ratings (Gravestock & Gregor-Greenleaf, 2008; Yao & Grady, 2005). To further undermine the use of summative evaluation for teaching improvement, faculty have doubts about the validity and reliability of the measurement instruments used for SRI because of diverse interpretations of effective teaching (Spooren et al., 2013). While instructors may be willing to accept students' feedback, at the same time they hold negative attitudes towards the summative use of SRI (Spooren et al., 2013). Therefore, it is necessary to develop a separate teaching evaluation for formative purposes, which is the purpose of this study.

What is Teaching Effectiveness?

Teaching effectiveness is not a new concept in higher education. However, there is no commonly agreed-upon definition or universal criteria to answer what effective teaching is (Benton & Cashin, 2012; Lehrer-Knafo, 2019; Shevlin, Banyard et al., 2000). When thinking about teaching effectiveness, it is necessary to consider the desired goals of teaching and learning in different contexts (Atkins & Brown, 2002). In other words, the meaning of *effective* in one context may not be the same in another (Atkins & Brown, 2002).

Nevertheless, the concept of teaching effectiveness is stakeholder relative. Students, teachers, and evaluator agencies may have different understandings of the meaning *effectiveness* (Fauth et al., 2020; Henard & Leprince-Ringuet, 2008). For example, the exemplary teachers may be concerned about effective teaching through lesson organization, lesson clarity, interests of the lesson, and positive classroom environment (Hativa et al., 2001). Delaney and colleagues (2010) conducted a study with 17,000 graduate and undergraduate students to explore the key factors that students perceived as essential for effective teaching. Students identified nine behavioral characteristics, including respectful, knowledgeable, approachable, engaging, communicative, organized, responsive, professional, and humorous (Delaney et al., 2010).

Recently, the focus of teaching effectiveness has been changed from observable teaching behaviors to students' learning. In theory, teaching quality should be conceptualized as a complex social process relying on the interactions between students and instructors (Fauth et al., 2020). Handelsman and colleagues (2007) demonstrated that "The instructor needs to consider what they want their students to know, understand, and be able to do and work back from there." Similarly, Hativa and colleagues (2001) defined effective teaching as the

"teaching that brings about effective and successful student learning that is deep and meaningful." In addition, Darling-Hammond and colleagues (2012) considered effective teaching as the instruction that enabled students to learn. Effective teaching should meet the demands of discipline, instructional goals, and students' needs in the teaching and learning environment (Darling-Hammond et al., 2012). "Effective teaching is about reaching achievement goals; it is about students learning what they are supposed to in a particular context, grade, or subject" (Berliner, 2005, p, 207). Carnell (2007) conducted a qualitative study with eight instructors teaching in higher education to examine university teachers' conceptions of effective teaching. Although the instructors have different teaching experiences, they all consider effective teaching to enable students' learning (Carnell, 2007).

Student Ratings of Instruction (SRI)

The use of SRI to measure and interpret teaching effectiveness has increased in higher education institutions since the 1900s. However, SRI has limitations on measuring teaching quality (Clayson & Haley, 2011; Pounder, 2007; Shevlin et al., 2000; Spooren et al., 2013; Uttl et al., 2017). First, both the content validity and construct validity of the commonly used teaching evaluation measurement instruments have been questioned (Spooren et al., 2013). Due to the lack of consensus of effective teachers' characteristics, there is a large variation in scope for the instruments used to measure teaching effectiveness, especially in the defined dimensions of teaching effectiveness (Spooren et al., 2013). In addition, the majority of measurement instruments used currently are designed by administrators without consideration of other essential stakeholders' views of effective teaching, which raises the question of content validity for the design of the instruments (Spooren et al., 2013).

Second, students' ratings can be affected by a variety of factors other than teaching practices, which raises the issues of accuracy of using SRI's results for high-stake decisions (Pounder, 2007; Shevlin et al., 2000; Worthington, 2002). Shevlin et al. (2000) conducted a study with 213 undergraduate students within a social science department at a UK University exploring the potential relationship between students' perception of the lecturer and their ratings for instruction. The results indicate that charisma factors account for 69% and 37% of the variation in lecturer ability and module attributes respectively (Shevlin et al., 2000). In addition, in a comprehensive review of the literature, Pounder (2007) synthesized a variety of potential student-level, course-level, and teacher-level factors that affected student ratings, concluding that relying only on SRI to measure teaching and learning in the classroom was problematic since the questions on the SRI failed to capture what happened in the classroom settings (Pounder, 2007). Worthington (2002) conducted a case study in a finance major course to investigate the effects of students' characteristics and their perceptions of the usage of SRI. The results suggest that the expected grade in the subject, student age, race, gender, and their perceptions of the evaluation process all have significant impacts on the ratings given to the instructors.

Third, SRI may discriminate instructors based on their background characteristics, especially for female instructors. Centra and Gaubatz (2000) conducted a study to investigate the relationship between students' gender and instructors' gender across classes regarding instruction ratings. The results indicate that both in the same class and across all classes, there is a significant difference between male and female students when rating female instructors, but no significant difference is detected for male instructions (Centra & Gaubatz, 2000). Female instructors are more likely to receive lower ratings from male students, even controlling for the effects of class size (Centra & Gaubatz, 2000).

Similarly, Young and colleagues (2009) conducted a study to explore the gender bias when rating the instructor and instructions, and the interaction between students' and instructors' characteristics, especially for the effects of gender. The results show a potential gender-preference while rating instructors on pedagogical characteristics and course content characteristics (Young et al., 2009). A more recent qualitative study conducted by Mitchell and Martin (2018) aimed to investigate the underlying relationship between gender and student evaluation of teaching. Based on analysis of the student comments, the authors found that students did evaluate their professors with gender bias. Students tend to comment on female instructors' appearance and personality more than male instructors and show less professional respect to woman instructors (Mitchell & Martin, 2018).

Considering the limitation of traditional SRI, the purpose of this study is to develop and validate a formative evaluation measurement instrument titled *Middle Semester Classroom Survey of Student Engagement and Learning* (MS-CSSEL) that can be used in college teaching for instructional improvement through the lens of students' engagement and learning gains. The measures of engagement and learning gains intend to provide comprehensive information for instructors to adjust teaching strategies during the course of instruction.

Theoretical Framework

Reliability and validity are two essential psychometric properties for any measurement instrument to be used. According to the *Standards for*

Educational and Psychological Testing (Joint Committee, 2014), validity is defined as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests" (p. 11). The concept of validity has evolved during the past century, from considering criterion validity to content validity (Strauss & Smith, 2009). Around 1980, construct validity was developed and accepted by scholars (Strauss & Smith, 2009). Today, it is commonly accepted that validity claims can be established based on content-related evidence, alignment of items with the defined theory, internal structure of items, response process evidence, criterion-related evidence, and consequence-related evidence.

Reliability refers to the "consistency of scores across replications of a testing procedure, regardless of how this consistency is estimated or reported" (Joint Committee, 2014). Reliability has been evaluated by a variety of coefficients depending on the measurement model being used, such as the Cronbach's Alpha reliability coefficient, generalizability coefficient, and Item Response Theory (IRT) information functions (Joint Committee, 2014).

In this study, we employed Rasch modeling to validate the psychometric properties of the newly developed measurement instrument. The item dimensionality estimates, the correlation between dimensions, fit statistics, item-person maps, and threshold estimates were generated to make claims about the instrument's construct validity. The Expected A Posteriori (EAP) reliability can be calculated to represent the degree of consistency for the instrument.

The research questions of this study are:

1) What evidence supports the multidimensional construct assumption of student engagement and learning gains?

2) Does the MS-CSSEL survey produce valid and reliable measures to assess student classroom engagement and learning gains?

Method

Purpose and Population of the Measurement Instrument

The purpose of the MS-CSSEL survey is to measure college student engagement and learning gains in the middle of the semester. The measures of engagement and learning gains will be used to infer whether the current instructions are effective, and for the instructors to adjust teaching for the rest of the course. This survey's target population is college students, and the setting for this survey is face-to-face college classrooms.

Student Engagement

While the conceptualizations of student engagement are diverse among researchers, there is an agreement that student engagement is a multidimensional concept with three coherent dimensions: behavioral, cognitive, and affective (Appleton et al., 2006; Fredricks et al., 2004; Kahu, 2013). The definition of behavioral engagement primarily relies on the idea of active involvement in academic, social, and extracurricular activities and the absence of negative behaviors for accomplishing positive learning outcomes (Fredricks et al., 2004; Fredricks & McColskey, 2012; Trowler, 2010). According to Kahu's (2013) framework of engagement, there were three specific subscales attached to behavioral engagement, time and effort allocated to educational activities, interactions with peers and instructors for educational purposes, and the extent of participation in learning activities (see Figure 1).

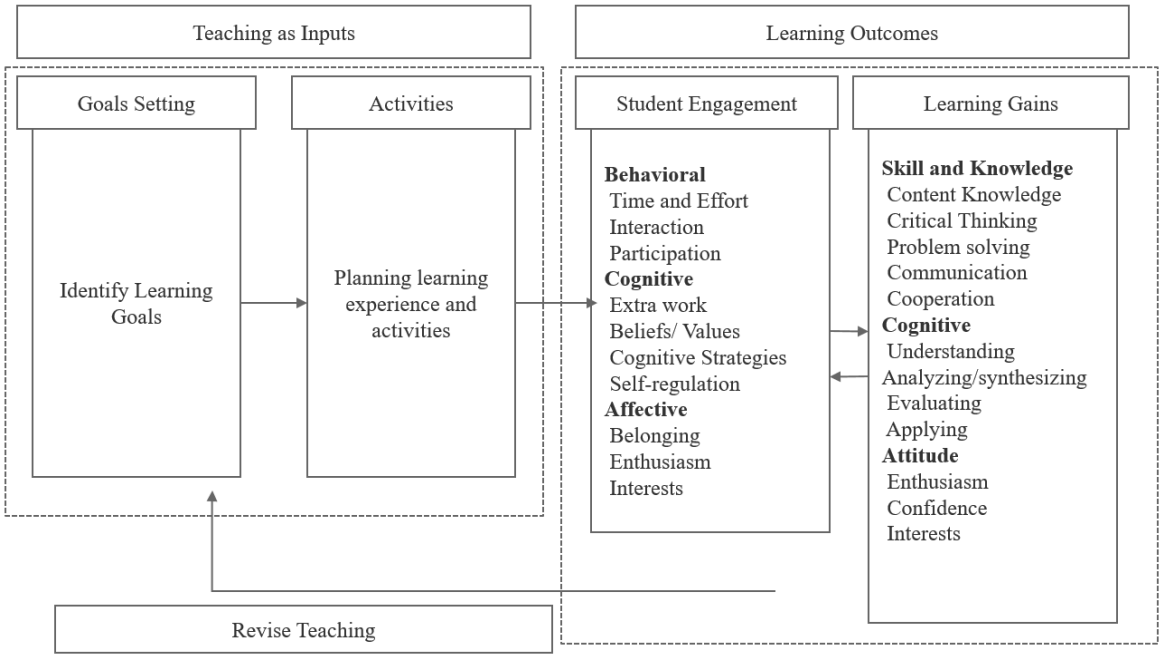
Cognitive engagement incorporates the idea of willingness to invest effort beyond the requirements for the course to understand complicated concepts and master skills (Fredricks et al., 2004; Fredricks & McColskey, 2012; Trowler, 2010). Cognitive engagement includes many latent factors that cannot be observed directly in the classroom such as self-regulation, values and beliefs about learning, cognitive and metacognitive strategies for learning (e.g., memorizing, synthesizing, understanding, evaluating, etc.), and personal goals and autonomy (Appleton et al., 2006; Fredricks et al., 2004; see Figure 1).

Emotional engagement focuses on the affective reactions to teachers, classmates, academics, and the institution, which impact the willingness to participate in school work (Fredricks et al., 2004; Fredricks & McColskey, 2012; Trowler, 2010). Kahu (2013) considered emotional engagement as a kind of attachment to schools or classes, while others considered it as enjoyment or interest in the learning activities. This study accepted the conceptual framework of Kahu (2013), which decomposed emotional engagement as enthusiasm for the courses, interest in the courses, and the sense of belonging to the classes (see Figure 1).

Learning Gains

After conducting a systematic literature review of learning gains and how they were measured, Rogaten et al. (2019) categorized learning gains into three different types following the commonly used ABC (i.e., Affective, Behavioral, and Cognitive) classification. Behavioral learning gains refer to skills, including study skills, leadership skills, team-work skills, and critical thinking skills (Rogaten et al., 2019). Cognitive learning gains are defined as understanding and cognitive

Figure 1
Conceptual Framework



abilities, such as analyzing, memorizing, synthesizing, evaluating, and applying abilities (Rogaten et al., 2019). Affective learning gains are mainly measured as the change of attitude during a course, such as confidence, motivation, and interests (Rogaten et al., 2019). In addition, how much students have learned about content knowledge is also an important indicator to measure students' learning gains. For this study, questions about content knowledge are incorporated into the behavioral dimension. Figure 1 presents the conceptual framework of the constructs measured in this study.

Instrument Construction

Based on Kahu's (2013) framework, the MS-CSSEL survey utilized five-point Likert-scale questions to assess student engagement. Behavioral engagement was assessed using 11 questions relating to how much time and effort students allocated to educational tasks, interactions with peers and instructors both inside and outside the classrooms, and attendance and participation. Cognitive engagement was assessed using 14 questions relating to how much extra work students did for the course, beliefs and values related to the course, cognitive strategies used to study the course content, and self-regulation strategies. Affective engagement was assessed using six items relating to sense of belonging in the course, enthusiasm, and interests in the subject.

Regarding the items measuring student learning gains, learning gains in skills and knowledge were assessed using seven questions that covered gains in content knowledge, critical thinking, problem-solving, communication, and cooperation. Cognitive learning gains were assessed using six questions related to understanding, analyzing and synthesizing, evaluating, and applying the course contents. Learning gains in attitude were assessed using six questions relating to enthusiasm, interests, and confidence of the content and the course.

In addition to the 50 items above, the MS-CSSEL survey includes two set of questions to ask to what extent a student's skills and knowledge gains, cognitive gains, and affective gains have been affected by some instructional teaching practices, such as the lecture, assigned class activities, graded assignments, etc.

At the end of the engagement section, two open-ended questions asked students how they had participated and engaged in this course and how the instructor could maintain and improve their engagement for the rest of the semester. Another two open-ended questions were attached to the learning gains section, which let students to describe what they had learned so far and provide suggestions to the instructor for the rest of the semester (see Appendix for the entire survey).

Data Collection

The MS-CSSEL survey was conducted to a target audience at one research University. In total, 634

undergraduate students in two introductory biology courses participated in this study (see Table 1). The MS-

Table 1
Descriptive Statistics of the Sample (N = 634)

	Category	N
Gender	Male	211
	Female	391
	Missing	32
Classification	Freshmen	424
	Sophomore	104
	Junior	39
	Senior	36
	Missing	31
Race	White	275
	Hispanic/Latino	43
	African American	54
	Native American	0
	Asian	189
	Others	40
	Missing	33

CSSEL survey was administrated in the middle of the Spring 2019 semester via a commercial online survey platform, SelectSurvey, which is similar to common platforms such as SurveyMonkey and Qualtrics. The online survey platform collected students' responses and generated descriptive report automatically. Students were invited by email with a survey link and they were given two weeks to complete the survey. The survey was conducted anonymously. The instructor provided 5 extra credit points (0.75% of the final grade) to the students for encouraging their participation in the survey. In total, 713 students were invited by email and 634 of them completed the survey with a response rate of 88.9%. As recommended by Wright and Douglas (1975), in order to obtain stable item calibration within ± ½ logit with 95% confidence interval for the errors, the minimum sample size is between 64 and 144. Thus, it is reasonable to assume that this study has an appropriate sample size to create a stable item calibration.

Traditionally, many Likert-scale evaluation instruments used in higher education have been developed according to the Classical Test Theory (CTT), which assumes that all items on the survey have the same standard error, and threshold estimates between categories for all items are equal (Van Zile-Tamsen, 2017). Researchers have demonstrated limitations when using CTT to analyze rating-scale or Likert-scale data (Bode & Wright, 1999).

Based on the validity and reliability theory, "using Rasch models to develop measurement instruments or Rasch modeling, is a systematic process in which items are purposefully constructed according to theory and empirically tested through Rasch models in order to produce a set of items that define a linear measurement scale" (Liu, 2020, p. 34). In order to address the limitation of CTT, this study applied the Rasch modeling approach to provide more accurate statistical evidence for the reliability and validity claim of the MS-CSSEL survey.

According to the literature, student engagement and learning gains are two multidimensional constructs. Thus, this study applied the multidimensional rating scale Rasch model for data analysis. All the items that measure classroom engagement adopted the same Likert-scaled categories from *strongly agree* to *strongly disagree*. Because all statements of items are in positive tones, *strongly agree* was coded as 5 and *strongly disagree* was coded as 1. The items that measured learning gains adopted a 5-point rating scale from 5 to 1. Items measuring students' classroom engagement and learning gains were analyzed separately following the same modeling procedures.

We used the "TAM" (Robitzsch et al., 2019) and "WrightMap" packages (Torres et al., 2014) in RStudio (RStudio Team, 2018) to do the Rasch analysis. Item fit statistics, EAP reliability coefficients, and item-person maps were generated. Additionally, we used Winstep 4.5.4 (Linacre, 2020) to test the appropriateness of the category structure. The item probability curves for student engagement and learning gains were drawn separately.

Results

Dimension Structure

Although student classroom engagement and learning gains are widely believed to be multidimensional constructs, there is little statistical evidence to prove the dimensional structure of these concepts. For testing the necessity of using multidimensional models, we ran both unidimensional and 3-dimensional Rating-scale Rasch models, and then compared the results. Because the multidimensional model hierarchically subsumes to the unidimensional models, the two models can be compared by testing the significant change in their deviance that describes the difference between the estimated model and the true model of the concept (Baghaei, 2012). Briggs and Wilson (2003) indicated that the difference of deviance between two estimated models was nearly a chi-square distribution with a degree of freedom of the difference between the number of parameters estimated in the two models. Janssen and De Boeck (1999) recommended selecting the model with significantly smaller deviance compared to estimated models. As shown in Table 2, the 3-dimensional Rasch model had significantly smaller deviance than the unidimensional model regarding

Table 2
Model Comparison

	Deviance	Npars	Chi-square	df	p
Engagement					
1-dimension	47697.30	35	570.76	5	<.05
3-dimension	47126.54	40			
Learning gains					
1-dimension	25527.94	23	339.20	5	<.05
3-dimension	25188.75	28			

Table 3
Correlation Matrix for the Dimensions of Classroom Engagement

	Behavioral	Cognitive	Affective
Behavioral	1		
Cognitive	0.78	1	
Affective	0.71	0.94	1

Table 4
Correlation Matrix for the Dimensions of Learning Gains

	Skills and Knowledge	Cognitive	Attitude
Skills and Knowledge	1		
Cognitive	0.95	1	
Attitude	0.84	0.88	1

student engagement, χ^2 (5) = 570.76, p < .05, suggesting that the 3-dimensional Rasch model was a better solution to model student engagement than the unidimensional model.

The same modeling process was applied to test the difference of deviance between the unidimensional model and 3-dimensional model used to analyze learning gains. The results indicated that the 3-dimensional Rasch modeling approach had a better fit to the true model of learning gains than a unidimensional model with a significant change of deviance, χ^2 (5) = 339.20, p < .05. Thus, we selected 3-dimensional Rasch model for testing the quality of the measures of engagement and learning gains.

In addition to the comparison of deviance, the correlations between dimensions of student engagement and learning gains were used to provide additional information for the preciseness of the measurement instrument (see Table 3 and Table 4)Table 3 and Table 4). When the multidimensional approach is used, "the higher the correlations, the greater the number of latent traits, and the shorter the subtests" will improves measurement precision significantly (Wang et al., 2004, p, 125). As shown in Table 5, the results indicated that students' behavioral engagement had a significantly high correlation to students' cognitive engagement (r = 0.78) and affective engagement (r = 0.71). In addition, students' cognitive engagement had a high-level correlation to students' affective engagement (r = 0.94). Although the smaller correlation estimates always help to differentiate dimensions, a higher correlation estimate does not necessarily imply an identical dimension (Baghaei, 2012). Table 6 presents the correlation matrix for learning gains. Overall, the dimensions of learning gains show a high degree of correlation (i.e., ranging from 0.84 to 0.95), which suggests that the measurement instrument has a high degree of precision.

Item Fit Statistics

Rasch modeling approach provides four indices for determining how the data fits the expected Rasch model. The mean square fit statistics (MNSQs) indicates how much the misfit observed between the Rasch model's expected item performance and the actual performance according to the data matrix (Bond & Fox, 2015). For the mean squared statistics, the closer to 1, the better the model-data-fit performed. For Likert-scale and rating scale questions, the commonly accepted range of the mean-square statistics is from 0.6 to 1.4 logits (Bond & Fox, 2015; Linacre, 2019). The standardized fit statistics (ZSTDs) indicate how likely the degree of misfit

Table 5
Fit Statistics for Classroom Engagement

	Outfit Statistics			Infit Statistics		
	MNSQ	ZSTD	p	MNSQ	ZSTD	p
Behavioral						
Item 1	1.25	4.55	0.00	1.23	4.35	0.00
Item 2	1.32	5.56	0.00	1.31	5.49	0.00
Item 3*	1.42	6.04	0.00	1.48	6.78	0.00
Item 4	0.96	-0.76	0.45	0.95	-0.89	0.37
Item 5*	1.54	9.03	0.00	1.53	8.91	0.00
Item 6	1.03	0.68	0.50	1.04	0.82	0.41
Item 7	0.97	-0.41	0.68	1.00	0.00	1.00
Item 8	0.82	-3.43	0.00	0.84	-3.18	0.00
Item 9	0.85	-3.21	0.00	0.84	-3.39	0.00
Item 10	1.37	6.09	0.00	1.37	6.22	0.00
Item 11*	2.93	23.58	0.00	2.92	23.46	0.00
Cognitive						
Item 12	1.23	3.81	0.00	1.19	3.21	0.00
Item 13	1.26	4.42	0.00	1.21	3.65	0.00
Item 14	1.08	1.32	0.19	1.13	2.07	0.04
Item 15	0.69	-5.70	0.00	0.72	-5.17	0.00
Item 16	0.80	-4.10	0.00	0.78	-4.52	0.00
Item 17	0.74	-5.43	0.00	0.69	-6.41	0.00
Item 18	0.78	-4.34	0.00	0.77	-4.54	0.00
Item 19	0.66	-6.72	0.00	0.67	-6.67	0.00
Item 20	0.63	-7.49	0.00	0.63	-7.61	0.00
Item 21	0.76	-4.61	0.00	0.74	-5.15	0.00
Item 22	0.74	-5.06	0.00	0.72	-5.29	0.00
Item 23	0.73	-5.17	0.00	0.73	-5.22	0.00
Item 24	0.70	-5.81	0.00	0.68	-6.19	0.00
Item 25	0.91	-1.56	0.12	0.90	-1.72	0.09
Affective						
Item 26	0.78	-4.22	0.00	0.78	-4.17	0.00
Item 27	0.75	-5.06	0.00	0.73	-5.45	0.00
Item 28	0.79	-3.93	0.00	0.78	-4.20	0.00
Item 29	0.65	-6.70	0.00	0.63	-7.22	0.00
Item 30	0.65	-6.50	0.00	0.69	-5.57	0.00
Item 31	0.83	-2.92	0.00	0.86	-2.40	0.02

expressed by mean square statistics will be observed (Bond & Fox, 2015). When the sample size is between 30 to 300, the acceptable range for the standardized fit statistic is from -.20 to 2.0 (Bond & Fox, 2015). Typically, the decision-making of model-data-fit depends on those four indices equally, but it is reasonable to make decisions according to some indices for a particular purpose (Bond & Fox, 2015). For example, if the sample size is larger than 300, the ZSTDs are more likely to be too sensitive (i.e., with many items failing to fit the model; Linacre, 2019). The standardized fit statistics highly depend on the sample size, which inflates putative Type I error rates; however, the mean square statistics are comparatively insensitive to sample size (Smith et al., 2008).

Over 600 students participated in this study; therefore, the decision-making of model-data-fit was made primarily based on the MNSQs. Overall, all 31

Table 6
Fit Statistics of Learning Gains

		Outfit Statistics			Infit Statistics		
		MNSQ	ZSTD	p	MNSQ	ZSTD	p
Skills and knowledge							
1	Item 32	0.87	-2.18	0.03	0.90	-1.66	0.10
2	Item 33	1.00	0.01	1.00	0.92	-1.28	0.20
3	Item 34	0.91	-1.62	0.11	0.95	-0.76	0.45
4	Item 35	0.78	-4.09	0.00	0.82	-3.29	0.00
5	Item 36	1.28	4.57	0.00	1.30	4.85	0.00
6	Item 37	1.06	0.99	0.32	1.06	1.05	0.29
7	Item 38	1.54*	7.83	0.00	1.60*	8.54	0.00
Cognitive							
8	Item 39	1.37	5.94	0.00	1.37	5.90	0.00
9	Item 40	0.70	-5.52	0.00	0.73	-4.79	0.00
10	Item 41	0.65	-6.60	0.00	0.68	-5.94	0.00
11	Item 42	0.72	-5.06	0.00	0.74	-4.67	0.00
12	Item 43	0.69	-5.81	0.00	0.72	-5.11	0.00
13	Item 44	0.87	-2.14	0.03	0.90	-1.74	0.08
Attitude							
14	Item 45	0.95	-0.76	0.44	1.02	0.39	0.70
15	Item 46	1.12	1.99	0.05	1.15	2.33	0.02
16	Item 47	0.84	-2.77	0.01	0.93	-1.24	0.21
17	Item 48	0.88	-2.13	0.03	0.88	-2.08	0.04
18	Item 49	0.71	-5.48	0.00	0.73	-5.02	0.00
19	Item 50	1.70*	9.47	0.00	1.70*	9.48	0.00

items measuring student classroom engagement fit the expected Rasch model well (Table 2). For classroom behavioral engagement, 8 out of 11 items demonstrated a good model-data-fit with a range of the MNSQs from 0.82 to 1.37. The MNSQs of items 3, 5, and 11 were outside the acceptable range (see Table 2). For all 14 items measuring cognitive engagement, MNSQs ranged from 0.63 to 1.23, which indicated an acceptable level of model-data-fit. All six items measuring student affective engagement also fit the expected Rasch model, with MNSQs ranging from 0.63 to 0.86.

As presented in Table 3, 6 out of 7 items measuring learning gains in skills and knowledge (i.e., items 32 to 38) had good model-data-fit, with MNSQs ranging from 0.78 to 1.30. Item seven, which asked how much students learned to communicate and work with peers to improve their learning, had more misinformation (Outfit MNSQ = 1.54; Infit MNSQ = 1.60). All items that measured learning gains in cognition (items 39-44) had acceptable MNSQs, ranging from 0.65 to 1.37. Five out of the six items that measured learning gains in attitude (items 45-50) fit the expected model well, with the MNSQs ranging from 0.84 to 1.15. Item 50, which asked whether students were willing to seek help from others when necessary, failed to fit the model, based both on Infit and Outfit MNSQ.

Internal Structure of Items

The item-person map, also called "Wright map," puts the person and item estimates in a same logit scale with the person ability estimates distributing on the left and the item difficulty estimates on the right. Generally, a good measurement instrument should be able to match sample's ability distribution with items' difficulty distribution (Liu, 2020, p. 40). In this study, person ability estimates represent levels of student engagement and extent of learning gains. On the Wright map, items were arranged by difficulty estimates from easier to agree with at the bottom and the harder to agree with on the top of the map. Individuals were arranged based on levels of engagement and learning gains from higher at the top to lower at the bottom.

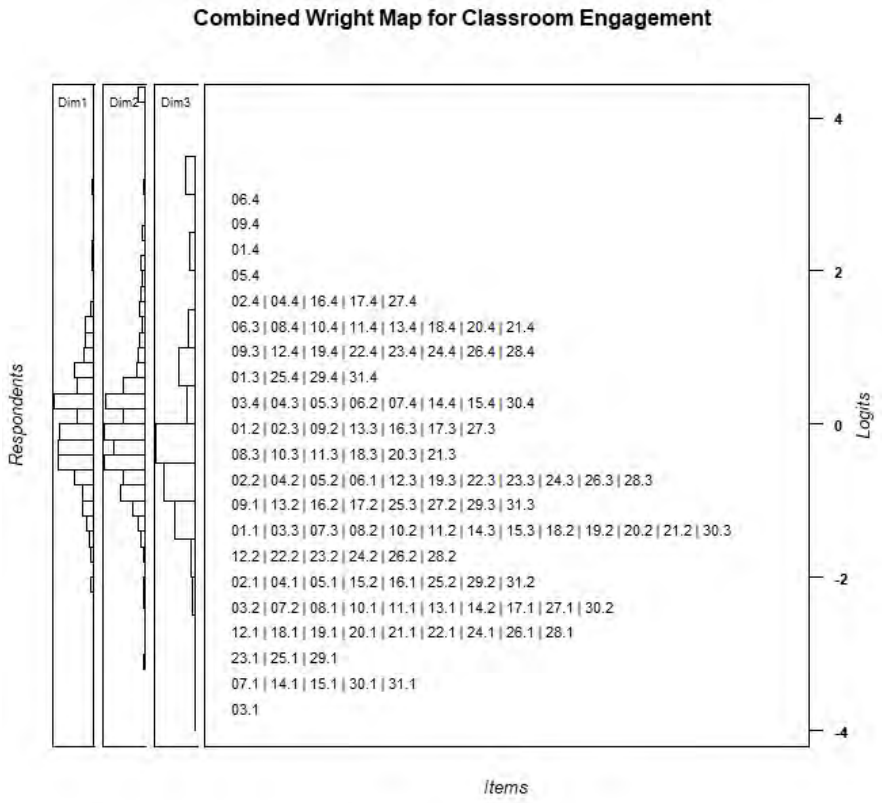
Evidence from the combined Wright Map for student classroom engagement (see Figure 2) indicated that levels of student classroom engagement in

behavioral, cognitive, and affective dimensions were normally distributed and the spread of student engagement estimates was sufficient with the logits ranging from -2.5 logits to 3.5. The mean logit for student engagement was slightly higher than the average of the item estimates suggesting that this set of questions was relatively easier for most students to agree with.

Separated Wright maps were produced for all dimensions of student engagement. Student behavioral engagement was determined from the time and efforts put into the course, interactions with peers and instructors, and participation in the course. The threshold Wright map (Figure 3) indicated that there were sufficient items to measure student classroom behavioral engagement. The results also indicated that the higher end of student engagement continued higher than the highest "difficult" item thresholds, which suggests that more items should be developed to assess the top end of student engagement precisely.

Figure 2

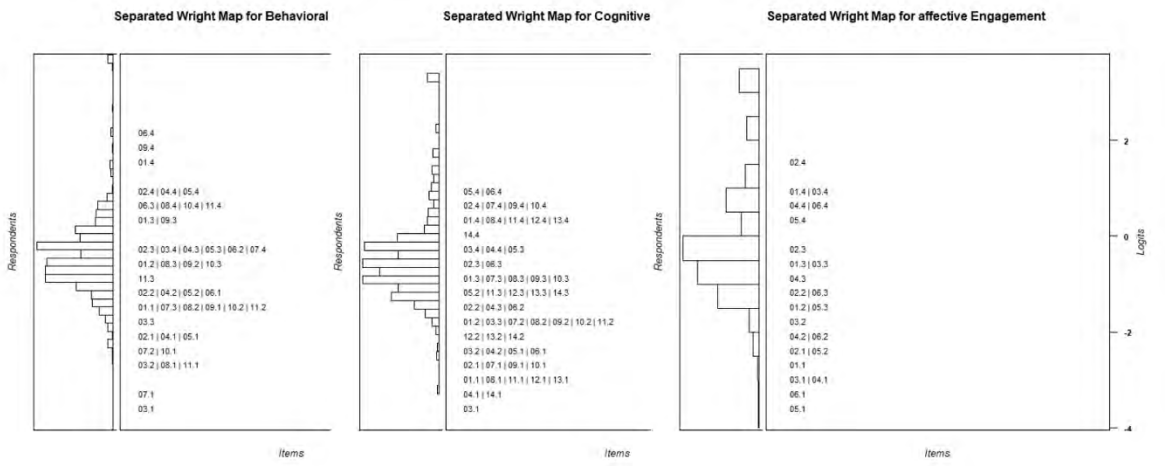
Combined Wright Map for Students' Classroom Engagement



As shown in Figure 3, overall, the distribution of cognitive engagement estimates was acceptable with a range of -2.0 logits to 4.5 logits. Overall, the threshold map indicated that the spread of items was sufficient to measure most levels of cognitive engagement. However, the higher end of cognitive engagement continued higher than the most "difficult" item, which suggests that more "difficult" questions were needed to capture higher cognitive engagement. Additionally, the efficiency of this set of questions was also not ideal, with 5 items (e.g., items 1,8,11,12,13) found to be at the same level of difficulty for this population. Thus, for further development, some of these questions should be combined or deleted in order to improve the efficiency of the survey.

As shown in Figure 3, the spread of students' affective engagement was acceptable with the range of affective engagement estimates ranging from -2.5 logits to 3.5 logits. The mean logit of the six items was slightly lower than the mean logit of students' affective engagement, which suggested that the items were relatively easier for students to agree with. The

Figure 3
Separated Wright Map for Behavioral, Cognitive, and affective Engagement



distribution of the threshold estimates had a sufficient spread to measure most levels of affective engagement, with the exception of the higher end. Thus, for further improvement, some more difficult questions should be considered for making this dimension more comprehensive.

For the items measuring student learning gains, it suggested that the spread of person estimates was acceptable ranging from -2.5 logits to 4 logits, and the shape of the distribution was nearly normal (see Figure 4). The mean logit of the items was slightly smaller than the average logit of students' learning gains, which

Figure 4
Combined Wright Map for Learning Goals

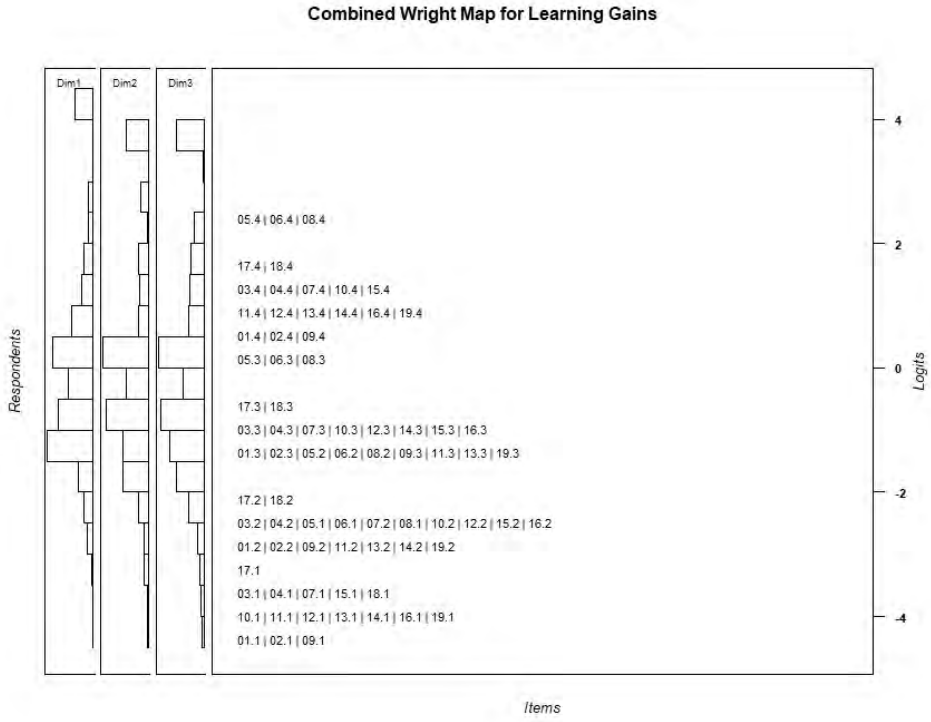
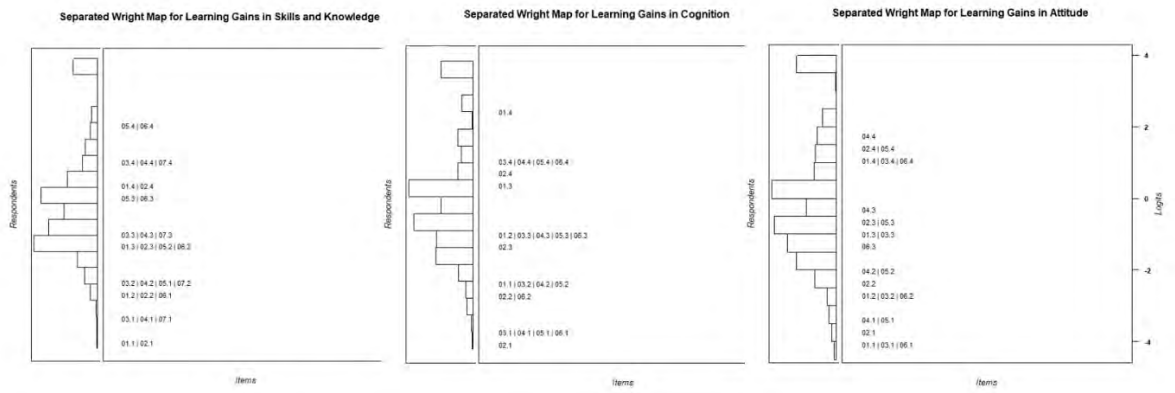


Figure 5
Separated wright Map for Learning Gains in Skills and Knowledge, Cognition, and Attitude



indicates that students gained more through classroom learning than what was measured by the 19 items together, in terms of the learning gains in skills and knowledge, cognition, and attitude.

The threshold map (Figure 5) of learning gains in skills and knowledge indicates that the separation of item difficulty is acceptable but not sufficient to measure all levels of student learning gains in skills and knowledge. Two gaps - between items 6 and 3, and between items 7 and 1 - indicate that more questions should be added.

The threshold map of learning gains in cognition (Figure 5) indicate that the six items measured some levels of cognitive learning gains, but the items were not sufficient to capture all levels of affective learning gains. The gap between items 1 and 3 suggests that more items should be added in this area.

Results from the threshold map for learning gains in attitude (Figure 5) indicate that the spread of the difficulty of threshold is acceptable, though there was a gap between items 4 and 6. Additionally, none of the questions successfully measured the highest and lowest levels of learning gains in attitude. This should also be addressed in future improvement of the survey.

Reliability

The Expected A Posteriori (EAP) Measures were calculated to test the reliability of the MS-CSSEL survey. As presented in Table 7, the results indicated a great extent of consistency regarding every sub-dimensions of classroom engagement and learning gains with the reliability coefficient ranging from 0.83 to 0.92. In particular, the reliability was 0.83 for behavioral engagement, 0.87 for cognitive engagement, and 0.83 for affective engagement. Regarding the section measuring students' learning gains, the reliability was 0.91 for learning gains in skills and knowledge, 0.92 for learning gains in cognition, and 0.90 for learning gains in attitude.

Table 7
Reliability of Engagement and Learning Gains

	EAP Reliability
Classroom Engagement	
Behavioral engagement	0.83
Cognitive engagement	0.87
Affective engagement	0.83
Learning gains	
Skills and knowledge	0.91
Cognitive	0.92
Attitude	0.90

Category Threshold Estimates

For the rating scale Rasch model, it is required that the average person estimates should advance monotonically from lower-level categories to higher, and the difference of neared threshold estimates should Advance by at least 1.4 logits for a 5-point scale (Bond & Fox, 2015).

Category Threshold Estimates

For the rating scale Rasch model, it is required that the average person estimates should advance monotonically from lower-level categories to higher, and the difference of neared threshold estimates should advance by at least 1.4 logits for a 5-point scale (Bond & Fox, 2015). The acceptable range of MNSQs for each category is from 0.6 to 1.4 (Bond & Fox, 2015). In this study, none of the MNSQs of categories were outside the acceptable range. The category Andrich threshold estimates suggested that the observed average for each category in both student engagement and learning gains

were increased monotonically. However, the difference of the Andrich threshold between "disagree" and "neutral" was 0.88, and the difference of the threshold between "neutral" and "agree" was 0.74, which suggested that students had difficulty distinguishing neutral with disagree and agree. As shown in Figure 6, the highest probability peak of "neutral" was less than .5, which indicated that this category was not functioning well. The results suggest that the response option neutral should be investigated further or deleted.

In terms of the category probability statistics for learning gains, as presented in Table 8, the majority of threshold estimates were acceptable. However, the difference between the second and third threshold estimates was 0.75. Because this is less than the minimum threshold of 1.0, it suggests that students might not be able to differentiate categories 2 and 3. As shown in Figure 7, the highest probability peak for the second category was less than 0.5, which suggested that this category was not functioning well. For further improvement, it is reasonable to consider using a 4-point rating-scale category structure for the items measuring learning gains.

Discussion

Validity and Reliability of the MS-CSSEL Survey

Overall, the Rasch analysis results suggest that the MS-CSSEL survey is a valid and reliable tool that can provide useful information to the instructor about student classroom engagement and learning gains in the middle of the semester. For the reliability of the MS-CSSEL survey, the EAP reliability estimates for each sub-dimension suggests that the survey has a large degree of consistency. In addition, the validity of the MS-CSSEL survey is supported by the following aspects: (a) this study supports the need to treat student engagement and learning gains as two multidimensional constructs by comparing the 3-dimensional model with the unidimensional model, (b) the moderately high correlation coefficient between sub-dimensions suggests a good internal relationship between dimensions, (c) the item fit statistics indicate that 47 out of 50 items contribute to the defined constructs, (d) the item-person maps show a good match between "person ability" and "item difficulty" for both the measures of classroom

Figure 6
Item Probability Curve for Student Engagement

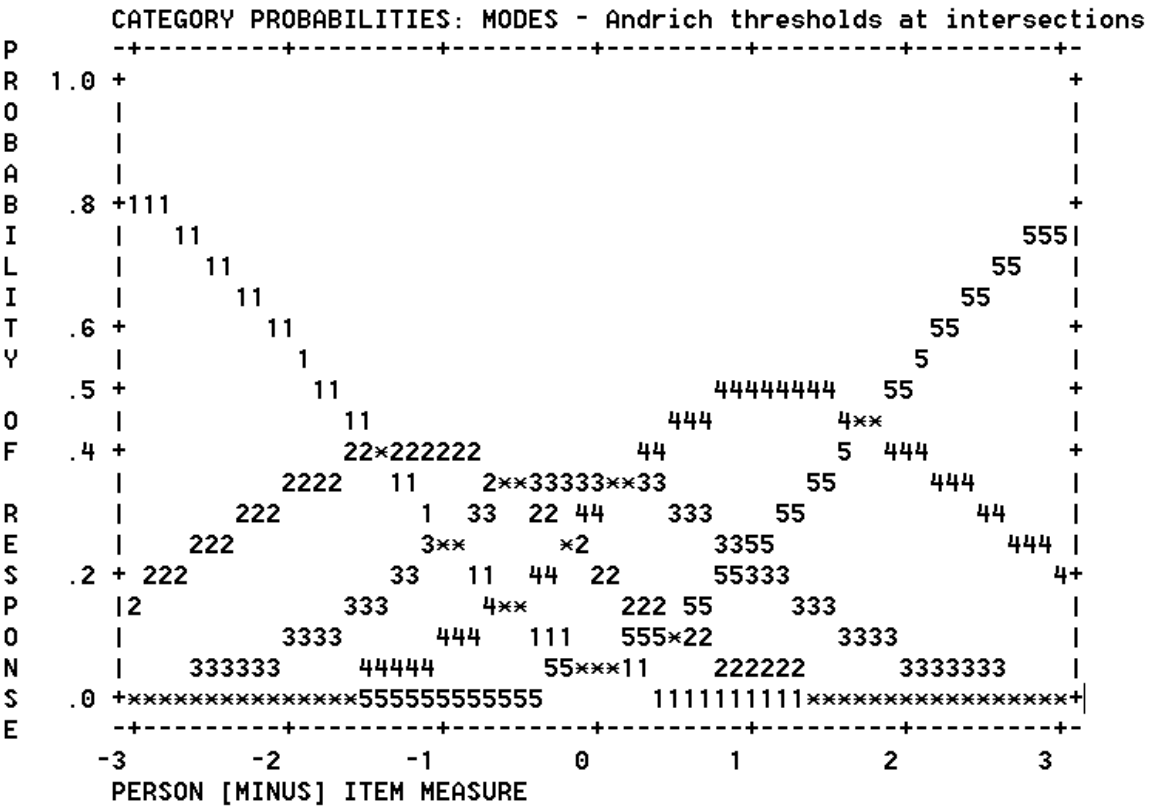
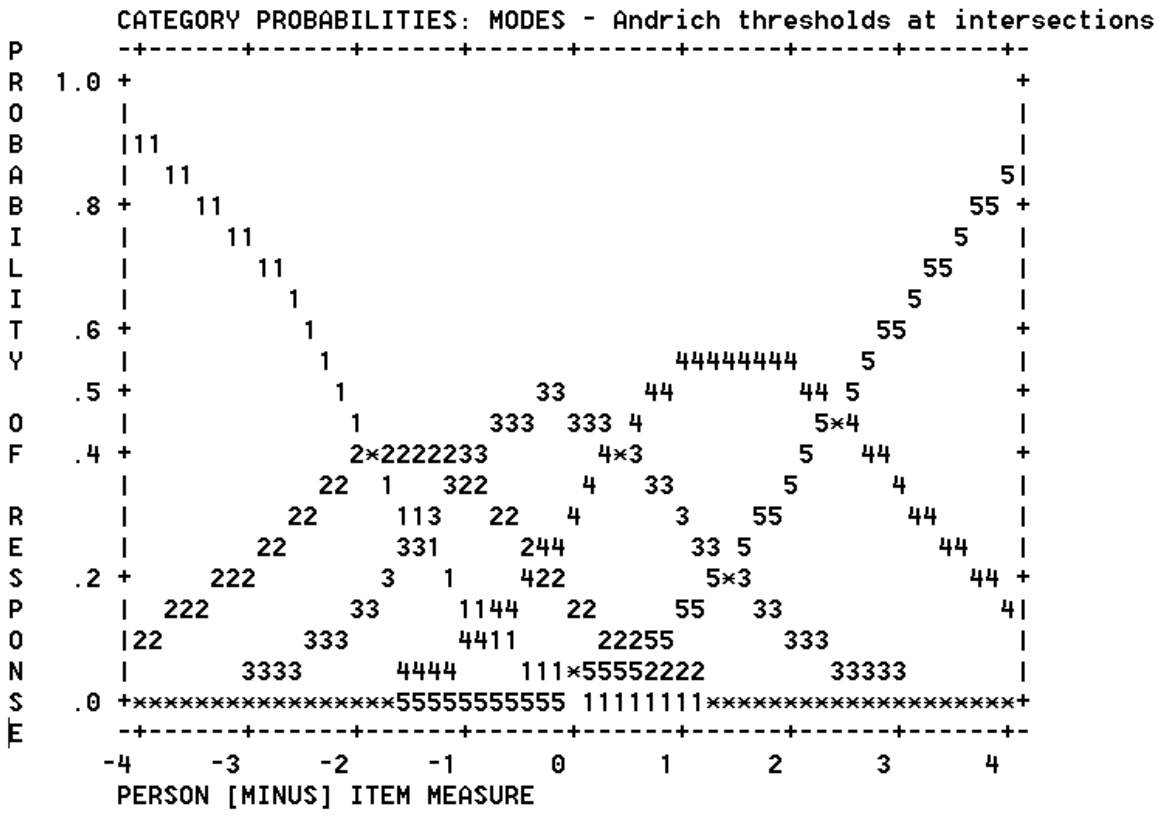


Table 8
Category Threshold Estimates

Category Label	MNSQs		Observed Average	Andrich Threshold
	Infit	Outfit		
Student engagement				
Strongly Disagree	1.17	1.30	-.55	None
Disagree	1.00	1.03	-.10	-1.41
Neutral	0.93	0.94	.36	-.53
Agree	0.89	0.86	.90	.21
Strongly Agree	1.03	1.02	1.52	1.73
Learning Gains				
1	1.22	1.36	-.89	None
2	1.00	1.06	-.35	-1.80
3	0.90	0.96	.34	-1.05
4	0.89	0.86	1.30	.43
5	1.06	1.04	2.60	2.42

Figure 7
Item Probability Curve for Learning Gains



engagement and learning gains, which suggests a good internal structure of items, and (e) the category threshold estimates for the measures of student engagement and learning gains indicate that each category adequately measured the constructs.

Although most items in the MS-CSSEL survey functioned as expected, a few items have been identified for further improvement. First, the question regarding student attendance showed poor model-data-fit. According to the definition of student behavioral engagement, attendance is an essential aspect that reflects how students engage in classroom learning. However, although the survey was conducted anonymously, students might not report their attendance accurately. Thus, this question should remain but needs revision. Second, the questions, "During class time, I regularly work with other students on work assigned by the instructor" and "Learning to communicate and work with peers to improve my learning" also showed misfit to the Rasch model. Those two questions are important aspects of engagement and learning gains. However, because most current large-enrollment university classes are lecture-based, collaborative learning between and among students occurs rarely. Thus, the instructor should consider how to incorporate collaborative learning in large classes or whether the items are meaningful to the course when using this survey. Third, the item-person map suggests that the items effectively measure students' engagement and learning gains, but some apparent gaps between items need to be taken into consideration for further improvement. Fourth, the results of category probability estimates implied that the response option "Neutral" could be deleted for the items measuring students' classroom engagement, and the response option "2" and "3" for the rating scale used for measuring learning gains could be collapsed. These items may be improved in continuous development and validation of the instrument.

Use of Results for Teaching Improvement

Instructors can use data produced by the MS-CSSEL survey for teaching improvement. The instructors do not need to conduct the statistical analysis reported in this paper, instead they can rely on each item's descriptive statistics to plan for teaching improvement. Any online survey platform, such as Survey Monkey, can generate descriptive statistics for each item automatically. The mean scores of students' responses to each question provide detailed information on students' average performance regarding each aspect of engagement and learning gains. Since the MS-CSSEL survey adopts a 5-point category structure, we recommend instructors pay attention to the items that have mean scores below 3. For example, the results of the behavioral engagement in this pilot study showed that students did not always ask questions in class (M=2.71) and had limited interaction with the instructor outside the classroom (M=2.25). Thus, for teaching improvement, the instructor may consider providing more opportunities for students to ask questions while teaching in the classroom and encourage them to interact with the instructor after class. Additionally, the information on how the designed instructional components affect students' learning gains can provide additional evidence for the instructor to consider further teaching improvement. For example, in this study, the average scores of "interactions with the instructor about learning" are lower than other instructional components in terms of the contributions to students' learning gains. Aligning with the findings identified through the Likert-type questions, the interaction between faculty and students is an essential aspect for further teaching improvement.

For researchers who will use the instrument, they should conduct Rasch analysis to obtain interval measures of student engagement and learning gains. First, the individual Rasch scores of engagement and learning gains of students will help understand the relationship between students' academic performance and their engagement and learning gains, which suggests better instructional strategies to teach students, especially for the low-performance students. For example, the individual "ability" scores could help researchers to reach the low-performing students to find out their strength and weakness in engagement and learning gains, and then to prepare more targeted mentoring and suggestions to the students for improving their academic performance in the rest of the course.

Second, by looking at the item-person map and the average scores of each item, researchers can identify in what aspects the majority of students' struggle. For example, the separated item-person map of behavioral engagement suggests that taking notes was the question that students most often agreed with, which meant that most students took notes during classroom learning. However, the item that asked whether students always discussed their learning process with the instructor outside the classroom was less commonly agreed with. Thus, to maximize students' engagement, the instructor should think about how to provide more opportunities for student-faculty interactions.

Finally, this survey can benefit students in terms of self-regulation and train them to be self-directed learners. By taking this survey, students will have an opportunity to monitor their engagement and reflect on what they have learned. Instructors can provide students the information on the average engagement and learning gains as well as how high-performance students engage in classroom learning as examples for other students to adjust their learning strategies. In this way, students will

have an opportunity to learn their strengths and weaknesses in the course.

Administration of the MS-CSSEL Survey

The MS-CSSEL survey is intended to measure student engagement and learning gains in the middle of the semester. Thus, it is appropriate to administer the survey around the middle of the semester in order to collect data for instructors to consider potential improvement while teaching the course.

The purpose of the MS-CSSEL survey is to help instructors identify the strengths and weaknesses of students' engagement and learning gains for adjusting teaching strategies. Thus, although the survey performed a high internal consistency and excellent construct validity, it should be used for formative evaluation of teaching instead of high-stakes faculty evaluation, according to the results. To protect students' privacy and encourage them to take the survey without concerns, we recommend administering the survey anonymously. Finally, because the survey is quite long, having students complete it on a purely voluntary basis will likely decrease the response rate. Thus, we recommend the instructor provide incentives to encourage student participation. For example, in this study, the instructor provided five extra points (~0.75% of the final grade) to the students who completed the survey on time. In this way, more than 88.9% of the students responded to the survey.

In conclusion, assessing teaching effectiveness through measuring student classroom engagement and learning gains is a viable way to address the lack of a unified definition of teaching effectiveness. Incorporating the MS-CSSEL survey in the middle of the semester will help instructors: (a) monitor and understand students' engagement and learning process, (b) gauge whether students' learning gains match with instructor's expectations, and (c) adjust instruction for the remainder of the course. Furthermore, it will prepare students to be self-directed learners.

References

American Educational Research Association, American Psychological Association, Joint Committee on Standards for Educational, Psychological Testing (US), & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. *American Educational Research Association*.

Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of School Psychology, 44*(5), 427-445.

Atkins, M., & Brown, G. (2002). *Effective teaching in higher education*. Routledge.

Baghaei, P. (2012). The application of multidimensional Rasch models in large scale assessment and validation: an empirical example. *Electronic Journal of Research in Educational Psychology*, *10*(1), 233–252.

Benton, S. L., & Cashin, W. E. (2012). Idea paper# 50 student ratings of teaching: A summary of research and literature.

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education, 17*(1), 48-62.

Berliner, D. (2005). The near impossibility of testing for teacher quality. *Journal of Teacher Education, 56*(3), 205 – 214.

Bode, R. K., & Wright, B. D. (1999). *In Higher education: Handbook of theory and research*. Springer.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*(1), 87-100.

Carnell, E. (2007). Conceptions of effective teaching in higher education: extending the boundaries. *Teaching in Higher Education, 12*(1), 25-40.

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education*, *71*(1), 17-33.

Chen, Y., & Hoshower, L. B. (2003). Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment & Evaluation in Higher Education, 28*(1), 71-88.

Clayson, D. E. (2013). Initial impressions and the student evaluation of teaching. *Journal of Education for Business, 88*(1), 26-35.

Clayson, D. E., & Haley, D. A. (2011). Are students telling us the truth? A critical look at the student evaluation of teaching. *Marketing Education Review, 21*(2), 101-112.

Darling-Hammond, L., Jaquith, A., & Hamilton, M. (2012). *Creating a comprehensive system for evaluating and supporting effective teaching*. Stanford Center for Opportunity Policy in Education.

Delaney, J. G., Johnson, A., Johnson, T. D., & Treslan, D. (2010). Students' perceptions of effective teaching in higher education. https://www.mun.ca/educ/faculty/mwatch/laura_treslan_SPETHE_Paper.pdf

Fauth, B. Göllner, R. Lenske, G. Praetorius, A. & Wagner, W. (2020). Who sees what? Conceptual

considerations on the measurement of teaching quality from different perspectives. *Zeitschrift Für Pädagogik, 66*, 138–155.

Fredricks, J. A., & McColskey, W. (2012). *Handbook of research on student engagement.* Springer.

Galbraith, C. S., Merrill, G. B., & Kline, D. M. (2012). Are student evaluations of teaching effectiveness valid for measuring student learning outcomes in business related classes? A neural network and bayesian analyses. *Research in Higher Education*, *53*(3), 353-374.

Gravestock, P., & Gregor-Greenleaf, E. (2008). *Student course evaluations: Research, models and trends*. Higher Education Quality Council of Ontario.

Handelsman, J., Miller, S., & Pfund, C. (2007). *Scientific teaching*. Macmillan.

Hativa, N., Barak, R., & Simhi, E. (2001). Exemplary university teachers: Knowledge and beliefs regarding effective teaching dimensions and strategies. *The Journal of Higher Education, 72*(6), 699-729.

Henard, F., & Leprince-Ringuet, S. (2008). *The path to quality teaching in higher education.* https://www.oecd.org/education/imhe/41692318.pdf

Janssen, R., & De Boeck, P. (1999). Confirmatory analyses of componential test structure using multidimensional item response theory. *Multivariate Behavioral Research*, *34*(2), 245-268.

Kahu, E. R. (2013). Framing student engagement in higher education. *Studies in Higher Education, 38*(5), 758-773.

Lehrer-Knafo, O. (2019). How to improve the quality of teaching in higher education? The application of the feedback conversation for the effectiveness of interpersonal communication. *EDUKACJA Quarterly*, *149*(2).

Linacre, J. M. (2020). Winsteps® Rasch measurement computer program. Winsteps.com

Liu, X. (2020). *Using and developing measurement instruments in science education: A Rasch Modeling Approach (*2nd ed.). IAP.

Mitchell, K. M., & Martin, J. (2018). Gender bias in student evaluations. *PS: Political Science & Politics, 51*(3), 648-652.

Nasser, F., & Fresko, B. (2002). Faculty views of student evaluation of college teaching. *Assessment & Evaluation in Higher Education, 27*(2), 187-198.

Pounder, J. (2007). Is student evaluation of teaching worthwhile? An analytical framework for answering the question. *Quality Assurance in Education*, *15*(2), 178–191.

Robitzsch, A., Kiefer, T., & Wu, M. (2019). *TAM: Test analysis modules. R package version 3.3-10.* https://CRAN.R-project.org/package=TAM

Rogaten, J., Rienties, B., Sharpe, R., Cross, S., Whitelock, D., Lygo-Baker, S., & Littlejohn, A. (2019). Reviewing affective, behavioural and cognitive learning gains in higher education. *Assessment & Evaluation in Higher Education, 44*(3), 321-337.

RStudio Team (2018). *RStudio: Integrated development for R. RStudio, Inc.* http://www.rstudio.com/.

Shevlin, M., Banyard, P., Davies, M., & Griffiths, M. (2000). The validity of student evaluation of teaching in higher education: love me, love my lectures? *Assessment & Evaluation in Higher Education, 25*(4), 397-405.

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology*, *8*(1), 33.

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.

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.

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology*, *5*, 1-25.

Torres Irribarra, D. & Freund, R. (2014). *Wright Map: IRT item-person map with ConQuest integration*. http://github.com/david-ti/wrightmap

Trowler, V. (2010). Student engagement literature review. *The Higher Education Academy,11*(1), 1-15.

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54*, 22-42.

Van Zile-Tamsen, C. (2017). Using Rasch analysis to inform rating scale development. *Research in Higher Education*, *58*(8), 922-933.

Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological methods*, *9*(1), 116-136.

Worthington, A. C. (2002). The impact of student perceptions and characteristics on teaching evaluations: a case study in finance education. *Assessment & Evaluation in Higher Education, 27*(1), 49-64.

Yao, Y., & Grady, M. L. (2005). How do faculty make formative use of student evaluation feedback? A multiple case study. *Journal of Personnel Evaluation in Education*, *18*(2), 107.

Young, S., Rush, L., & Shaw, D. (2009). Evaluating gender bias in ratings of university instructors' teaching effectiveness. *International Journal for the Scholarship of Teaching and Learning, 3*(2), n2.

Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education*, *12*(1), 55-76.

REN LIU is a Ph.D. candidate in the Department of Learning and Instruction, Graduate School of Education, University at Buffalo, State University of New York. He conducts research in the areas of college teaching improvement, measurement development using Rasch model, program evaluation, and application of Rasch measurement in higher education.

XIUFENG LIU is a Professor in the Department of Learning and Instruction, Graduate School of Education, University at Buffalo, State University of New York. He conducts research in closely related areas of measurement and evaluation of STEM education, applications of Rasch measurement, and STEM teaching and learning. He was the inaugural director of the Center for Educational Innovation with a mission for improving university teaching and student learning. Among the books he has published is *Using and Developing Measurement Instruments in Science Education: A Rasch Modeling Approach* (Second Edition, 2020, Information Age Publishing).

LARA HUTSON is a Clinical Associate Professor and Director of Undergraduate Studies in the Department of Biological Sciences at the University at Buffalo (SUNY). She is also coordinator and instructor of the second-semester introductory majors' biology course (Cell Biology) and teaches biochemistry. Dr. Hutson's most recent projects aim to reduce D/R/F/W rates in introductory STEM courses, including Attendance Tracking and Intervention (funded by the UB President's Circle); SUNY Excels, a collaboration between STEM instructors at the SUNY Centers (funded by the State of New York); and UB Excite, a course redesign program (also funded by the State of New York).

Appendix

Part I: Classroom engagement

Behavioral Engagement	**Strongly agree** ▼	**Agree** ▼	**Neutral** ▼	**Disagree** ▼	**Strongly disagree** ▼
1. I always ask content-related questions in class.	□	□	□	□	□
2. I always complete assigned readings before coming to class.	□	□	□	□	□
3. I **always** take notes during class.	□	□	□	□	□
4. I **always** review my notes of previous classes before coming to the next class.	□	□	□	□	□
5. During class time, I **regularly** work with other students on work assigned by the instructor.	□	□	□	□	□
6. I **always** discuss my learning progress (e.g., grades, assignments, learning difficulties) with the instructor out of classroom.	□	□	□	□	□
7. I **always** pay attention to the instruction during class time.	□	□	□	□	□
8. If I have a difficulty in understanding something, I always seek additional help.	□	□	□	□	□
9. I **always** contribute to class discussion in class.	□	□	□	□	□
10. I **always** discuss assignments with other students.	□	□	□	□	□
11. I have been absent in class fewer than 2 times so far.	□	□	□	□	□
Cognitive Engagement					
12. I spend more time and effort on this course (e.g., assignments, studying, reviewing notes) than on other courses.	□	□	□	□	□
13. I spend much time and effort in finding additional resources to help me complete the course work.	□	□	□	□	□
14. I clearly understand the value and the importance of this course for my future learning and career.	□	□	□	□	□
15. I work hard in order to meet the instructor's expectation.	□	□	□	□	□
16. I fully understand the course content.	□	□	□	□	□
17. I memorize the course content (e.g., definitions, facts, ideas, or methods) and can recall them well.	□	□	□	□	□
18. I always try to decompose an idea or theory to identify its components or elements.	□	□	□	□	□

19. I synthesize knowledge (e.g., ideas, information, experiences) into more comprehensive interpretations and relationships.	□	□	□	□	□
20. I examine critically in order to make judgment about the value of information, arguments, or methods learned from this course.	□	□	□	□	□
21. I apply new knowledge and skills learned in this course to solve practical problems.	□	□	□	□	□
22. Before beginning a task, I plan for appropriate strategies and allocate sufficient time.	□	□	□	□	□
23. I monitor my learning progress during the course.	□	□	□	□	□
24. I consider the instructor's feedback of my learning performance carefully and adjust my learning accordingly.	□	□	□	□	□
25. I fully understand my strength and weakness of learning in this course.	□	□	□	□	□
Affective Engagement					
26. I feel I belong to this course as a learning community.	□	□	□	□	□
27. I feel I have a voice in this course.	□	□	□	□	□
28. I feel comfortable to talk to the instructor.	□	□	□	□	□
29. I feel supported by the instructor.	□	□	□	□	□
30. I am enthusiastic about learning new things.	□	□	□	□	□
31. I am interested in the course content.	□	□	□	□	□

1. Please describe how you have participated and engaged in this course so far.
2. Please suggest how the instructor may maintain and further improve your participation and engagement during the rest of the course.

Part II: Learning gains: How much have you learned so far in this course

Please rate your learning gains in the following aspects from 1(lowest) to 5(highest)

Skills and Knowledge	**1** ▼	**2** ▼	**3** ▼	**4** ▼	**5** ▼
32. Gaining factual knowledge (terminology, classifications, methods, trends).	□	□	□	□	□
33. Learning fundamental principles, generalizations or theories.	□	□	□	□	□
34. Learning how to find and use resources for answering questions or solving problems.	□	□	□	□	□
35. Developing specific skills, competencies and points of view needed by professionals in the field most closely related to this course.	□	□	□	□	□
36. Developing skill in expressing yourself orally or in writing.	□	□	□	□	□
37. Learning to communicate with the instructor to improve my learning.	□	□	□	□	□
38. Learning to communicate and work with peers to improve my learning.	□	□	□	□	□
Cognitive					
39. Developing creative capacities (writing, inventing, designing).	□	□	□	□	□
40. Gaining a broader understanding and appreciation of key concepts of this course.	□	□	□	□	□
41. Learning to analyze and critically evaluate ideas, arguments and points of view related to the key topics in this course.	□	□	□	□	□
42. Learning to synthesize and organize new knowledge into a more complex and comprehensive way.	□	□	□	□	□
43. Learning to apply course material (to improve thinking, problem solving and decision making)	□	□	□	□	□
44. Learning to apply ideas from this class to ideas encountered in other classes within this subject area.	□	□	□	□	□
In general, how much has each of the following aspects of the course helped your cognitive learning gains (Q32-Q45)?					
• Lecturing presentation	□	□	□	□	□
• Assigned class activities (e.g., discussions, problem solving, case studies)	□	□	□	□	□
• Graded assignments	□	□	□	□	□
• Feedback on my work of class assignments and exams from instructors	□	□	□	□	□
• Course materials (e.g., textbooks and supplementary readings)	□	□	□	□	□
• Examinations and quizzes	□	□	□	□	□

• Online notes or resources posted by instructor	□	□	□	□	□
• Interaction with the instructor about your learning	□	□	□	□	□
Attitude					
45. Acquiring an interest in learning more knowledge and skills from this course, and having interests to take additional class in this field.	□	□	□	□	□
46. Developing a clearer understanding of, and commitment to, personal values.	□	□	□	□	□
47. Enthusiasm for the subject.	□	□	□	□	□
48. Confidence that you understand the course materials.	□	□	□	□	□
49. Feeling comfortable in working with complex ideas or tasks in this field.	□	□	□	□	□
50. Willing to seek help from others (e.g., teacher, peers, TA) when necessary.	□	□	□	□	□
In general, how much has each of the following aspects of the course helped your affective learning gains (Q45-Q50)?					
• Lecturing presentation	□	□	□	□	□
• Assigned class activities (e.g., discussions, problem solving, case studies)	□	□	□	□	□
• Graded assignments	□	□	□	□	□
• Feedback on my work of class assignments and exams from instructors	□	□	□	□	□
• Course materials (e.g., textbooks and supplementary readings)	□	□	□	□	□
• Examinations and quizzes	□	□	□	□	□
• Online notes or resources posted by instructor	□	□	□	□	□
• Interactions with the instructor about your learning	□	□	□	□	□

1. Please describe what you have learned in this course so far?
2. Please suggest how the instructor may maintain and further improve instruction in order to maximize your learning gain during the rest of the course?

Part III: Demographic

If you do not want to respond to any of the questions below, you can choose to skip it

What is your Gender	Male	□	Female	□
What is your Race? (please choose all that apply)	White	□	Hispanic/ Latino	□
	African American	□	Native American	□
	Asian	□	Others, please specify	
Are you a first-generation college student?	Yes	□	No	□
What is your classification?	Freshmen	□	Sophomore	□
	Junior	□	Senior	□
What are the reasons to register in this course?	Required	□	Selective	□
	Others	--------		