10-2023

Investigating Evidence in Support of Validity and Reliability for Data collected with the meaningful learning in the laboratory instrument (MLLI)

Elizabeth Vaughan
Portland State University

A. Montoya-Cowan
Portland State University

Jack Barbera
Portland State University, jbarbera@pdx.edu

Citation Details

Published as: Vaughan, E. B., Montoya-Cowan, A., & Barbera, J. (2024). Investigating evidence in support of validity and reliability for data collected with the meaningful learning in the laboratory instrument (MLLI). Chemistry Education Research and Practice, 25(1), 313-326.

Investigating Evidence in Support of Validity and Reliability for Data Collected with the Meaningful Learning in the Laboratory Instrument (MLLI)

Elizabeth B. Vaughan, Amanda Cochran, and Jack Barbera

Abstract

The Meaningful Learning in the Laboratory Instrument (MLLI) was designed to measure students' expectations before and after their laboratory courses and experiences. Although the MLLI has been used in various studies and laboratory environments to investigate students' cognitive and affective laboratory expectations, the authors of the instrument reported a discrepancy between the intended factor structure of the MLLI and the factor structure suggested by the data collected in preliminary studies. Therefore, the aim of this study was to investigate the validity and reliability evidence related to data collected with the MLLI, especially that related to structural validity. Evidence to support structural validity would provide greater meaning for the reporting and interpretation of MLLI scores. In this study, two possible *a priori* models for the factor structure of data collected from multiple institutions with the MLLI were investigated using confirmatory factory analysis (CFA). This initial investigation found poor data-model fit for each of the two tested models. Cognitive interviews and free response items were then used to inform modifications to the two *a priori* structures, and a third alternative structure, which included a negative method factor, was also investigated. Once a best fitting model was identified, further model revisions were informed by a combination of modification indices and qualitative data. Evidence of adequate-to-good data model fit was found for the final revised version of the MLLI, deemed the MLLIv2. Additionally, evidence of both internal structure validity and single administration reliability were found for each of the MLLIv2 factors. The structure of the data from these items leads to scale scores that likely represent student expectations that contribute to meaningful learning and student expectations that detract from meaningful learning. As the results of this study provide the first psychometrically supported scales for MLLI data, they have implications on the future reporting and analyses of MLLI scores.

Introduction

In recent years, there has been a great deal of interest in investigating and improving chemistry laboratory courses (Bretz, 2019). As a part of this trend, many chemistry education researchers and practitioners have begun moving away from confirmatory "cookbook" style laboratory activities toward inquiry-based and/or research-based laboratory activities (Weaver et al., 2008; Grushow et al., 2021). As laboratory courses are being transformed, it is important for researchers and practitioners alike to investigate the impact that these changes have on student learning experiences.

It has been noted in the chemistry education literature that students' laboratory expectations (i.e., students' perceptions about what they will learn, think, do, or feel in a laboratory course) likely have an impact on their laboratory learning experiences and behaviors (DeKorver & Towns, 2015; Galloway and Bretz, 2015a; Wang et al., 2021). More specifically, researchers suggest that increasing the alignment between student expectations and laboratory learning goals could improve student buy-in and engagement in these courses (DeKorver & Towns, 2015; Galloway & Bretz, 2015a; Rovers et al., 2018), potentially improving other student outcomes, such as course grades and persistence (Lester, 2013). One of the lenses through which students' laboratory expectations and experiences can be assessed is meaningful learning.

Meaningful Learning

Joseph Novak's Theory of Education states that "Meaningful learning underlies the constructive integration of thinking, feeling, and acting leading to empowerment for commitment and responsibility" (Novak, 2010). In the context of an undergraduate teaching laboratory, "how a student chooses to act (psychomotor) in the lab depends on how they think about (cognitive) and feel toward (affective) their laboratory experiences" (Galloway and Bretz, 2015a). Novak's Theory of Education was highly influenced by the work of David Ausubel, whose theories highlighted the distinction between rote and meaningful learning (Ausubel, 1962, 1963, 1968; Novak, 2003). Ausubel theorized that meaningful learning occurs when an effort is made by the learner to connect newly acquired information to knowledge that the learner already possesses. Rote learning, on the other hand, occurs when little to no effort is made by the learner to relate new information to existing knowledge (Ausubel, 1962, 1963, 1968; Novak, 2003). Furthermore,

Ausubel and Novak theorized that learning exists on a continuum "from extreme rote to highly meaningful, with key factors being the strength of the learner's commitment to learn meaningfully and the quantity and quality of organization of her/his relevant knowledge." (Novak, 2003).

Measuring Meaningful Learning

In the context of undergraduate chemistry laboratory courses, Novak's Theory of Meaningful Learning has been operationalized as the Meaningful Learning in the Laboratory Instrument (MLLI) (Galloway and Bretz, 2015a). This tool was designed to measure students' cognitive and affective learning expectations (pre-course) and experiences (post-course) in their undergraduate chemistry laboratory courses. Due to the inherent psychomotor nature of chemistry laboratory courses, the MLLI authors chose not to include any solely psychomotor items. Instead, they wanted to capture the extent to which "students integrate their thinking and feeling with the doing" (Galloway and Bretz, 2015a). The MLLI contains 30 items, each author-assigned into one of three categories: cognitive (16 items), affective (8 items), and cognitive/affective (6 items). The authors also identified that 16 of the items are positively worded, while the other 14 are negatively worded. Each of the MLLI items share a common item stem, which focuses the items on in class laboratory experiences directly related to performing experiments. For example, the pre-course item stem states: 'When performing experiments in my chemistry laboratory course this semester, I expect…'. Students are then asked to respond to a variety of statements, for example, '… to learn chemistry that will be useful in my life' on a 0 to 100 response scale, where students indicate their agreement (100%, completely agree) or disagreement (0%, completely disagree) to each item statement by moving a slider bar in 1% increments. (Galloway and Bretz, 2015a).

The MLLI authors investigated the factor structure of the data collected with their instrument through exploratory factor analysis (EFA). Their analysis suggested that the most appropriate structure of the data, for both the pre- and post-assessment data, was a two-factor structure in which one factor contained 13 negatively worded items, while the other factor contained 15 positively worded items. Additionally, one item did not strongly load on either factor and one item cross-loaded strongly on both factors. The authors suggested that the two factors accounted for "items contributing to meaningful learning" (positively worded items) and

"items inhibiting meaningful learning" (negatively worded items). (Galloway and Bretz, 2015a). This statistically supported two-factor structure did not match the three domains of item categorization intended by the MLLI authors. To date, there has been no published quantitative support for the proposed three domain categories of the MLLI.

MLLI Uses and Score Reporting

Since its publication, the Meaningful Learning in the Laboratory Instrument has been used to examine students' expectations and experiences in a variety of research studies and teaching laboratory environments. Throughout their work, the MLLI authors used the averages of the items assigned to the proposed categories (i.e., cognitive, affective, and cognitive/affective) to investigate the expectations and experiences of both general and organic chemistry students (Galloway and Bretz, 2015a, b, c). Various other researchers have also used the MLLI to investigate students' expectations and experiences when implementing changes in their laboratory courses. For example, a study published in 2017 used the MLLI to investigate the perspectives of students taking upper-division laboratory courses focused on "analytical measurements and physical measurements." In this study, the MLLI was one piece of a mixed methods approach investigating the utility of pre-laboratory video resources in rotational style courses (Schmidt-McCormack et al., 2017). To do so, MLLI data were calculated as the median scores for each student's responses to the cognitive, affective, and cognitive/affective item categories (Schmidt-McCormack et al., 2017). Additionally, studies published as recently as 2022 have used the MLLI and its three proposed categories in their investigations. For example, one study reported the average scores of the cognitive, affective, and cognitive/affective categories to assist in the investigation of the differences between traditional chemistry courses and integrated lecture-lab block courses (Lau et al., 2023).

Not all studies utilizing the MLLI report scores from the three proposed categories. In 2019, George-Williams and colleagues published a study investigating student expectations and experiences, along with staff perceptions of students' experiences in teaching laboratories, through the lens of meaningful learning. To facilitate pen-and-paper data collection, and to collect data from instructors as well as students, slightly modified versions of the MLLI items and response scales were used. In their publication, the researchers reported that "a factor analysis did not show factors that aligned with original ones raised by Galloway and Bretz

(2015a) (affective, cognitive and affective/cognitive)." For this reason, George-Williams and colleagues decided not to report scores for the affective, cognitive, and affective/cognitive item categories and instead compared participants' responses at the item level (2019). This methodological decision highlights an important aspect of instrument development and utilization: Before data can be meaningfully scored and interpreted by researchers, it is important to collect evidence supporting the validity and reliability of the data produced by a measure.

Validity and Reliability

Self-report surveys, like the MLLI, generally consist of a set of items that participants directly respond to. These items are theoretically related to an unobserved (latent) variable or variables (i.e., students' expectations related to meaningful learning). Before the data produced by the MLLI, or any other measure intended to investigate latent variables, can be scored and interpreted, evidence supporting the validity and reliability of the data must be gathered (Lewis, 2022; Stains, 2022). Evidence of validity provides support that an instrument measures what it is intended to measure, while evidence of reliability provides information about the consistency of the data (American Educational Research Association, 2014; Arjoon et al., 2013). There are various types of validity and reliability evidence that can be assessed during the development and/or usage of a given measure, including response process validity, internal structure validity, and single administration reliability.

Response Process Validity

Response process validity focuses on how participants interpret and respond to the items included in a measure (Arjoon et al., 2013; Deng et al., 2021, Collins, 2003). Typically collected through cognitive interviews and/or short-answer survey items, response process data can provide insight into respondents' thought processes when responding to items. Collecting this type of data allows researchers to explore the respondents' understanding of the nuances of each item. Evidence of response process validity is especially important when a measure is being developed, but it is also necessary when a measure is adapted or modified to a new environment, as participants in a new environment may not interpret items in the same way as respondents from the environment in which the measure was originally developed.

Internal Structure Validity

Validity evidence based on internal structure is concerned with the relations between items and latent constructs and how these relations match to the hypothetical structure of the construct (Arjoon et al., 2013, Worthington & Whittaker, 2006). This evidence can be supported through factor analysis, either exploratory factor analysis (EFA), where the relations are analyzed without a preconceived structure, or confirmatory factor analysis (CFA), where an estimate of how well the data fit an *a priori* model is obtained (Arjoon et al., 2013). Evidence of structural validity is important, both when a measure is developed, and when a measure is adapted or modified to a new environment. Before a participant's 'score' for a set of items representing a latent construct can be interpreted, evidence supporting the item grouping (i.e., factor structure) must be collected (Lewis, 2022; Stains, 2022).

Single Administration Reliability

Single-administration reliability is concerned with how consistent participants' responses are to items measuring the same construct. Currently in STEM education literature, one of the most commonly reported coefficients of reliability is Cronbach's alpha (Barbera et. al., 2020; Taber, 2017). However, as noted by Komperda, Pentecost, and Barbera in 2018, alpha should only be used for data models that have equal item loadings. Since most measures used in education research are not designed to meet this requirement, single-administration reliability can instead be estimated using McDonald's omega, which allows for unequal factor loadings and describes the amount of the observed variance explained by the construct divided by the total variance (Komperda et al., 2018). Collecting evidence of single-administration reliability provides researchers with information about the relations between individual items and a participant's 'score' for a set of items representing a latent construct.

Goals of This Study

Without appropriate evidence to support the validity and reliability of the data collected with the MLLI, the student responses gathered using the instrument cannot be meaningfully scored or interpreted. As part of a larger project to explore the relations among students' expectations, buy-in, and engagement in lower-division undergraduate chemistry laboratory courses, this study aims to investigate the validity and reliability evidence for the data collected with the MLLI in the populations under investigation. Therefore, the research questions guiding this study are as follows:

1. What evidence of validity and reliability supports interpreting data collected with the MLLI as measures of student expectations in lower division undergraduate laboratory courses?
2. If insufficient evidence is found, what modifications are supported by the data collected from this population?
3. If modifications are necessary, what evidence of validity and reliability supports interpreting data collected with the modified instrument?

Methods

All data collected within this study was approved by the Institutional Review Board (IRB) at Portland State University, and appropriate consent was obtained from students as required by the IRB.

Participants and Data Collection

Data collection for this study included the collection of qualitative response process validity data and the collection of quantitative data to assess evidence of structural validity and single administration reliability. Before any data were collected using the MLLI, the item stems were edited to remove the word 'semester'. This generalization was made so that data collection at institutions with different types of calendars (i.e., semesters vs. quarters) was identical. Therefore, the item stems used in this study were 'When performing experiments in my chemistry laboratory course, I expect…' and in the post-course assessment, 'When I performed experiments in my chemistry laboratory course, I…'.

Qualitative Data Collection

In this study, response process data were collected through both interviews and free response items. The sample for response process interviews consisted of Portland State University students taking a first-term general or organic chemistry laboratory course in the fall of 2020. Students' interest in participating in an interview and consent were collected via Qualtrics. Based on each student's availability, interviews were scheduled and conducted via Zoom. Each interview was approximately 45 minutes long. During each interview, a copy of the MLLI was provided to the participant and they were first directed to complete the items. Interviewees were then asked to read each MLLI item aloud, state which response (from 0 to

100) they selected, and explain why they selected that response value. Follow-up questions were asked, as needed, to gain more details about their understanding of the items and/or response reasoning. All interviews were audio-video recorded over Zoom and transcribed before analysis.

The sample for open-ended response process survey items consisted of Portland State University students taking a second-term general or organic chemistry laboratory course in the winter term of 2021. These data were collected through open-ended survey items distributed via Qualtrics in the first two weeks of each laboratory course. Survey recruitment consisted of a video announcement that was pre-recoded by the first-author (E.B.V.) and presented by the graduate teaching assistants in each laboratory section near the beginning of the first course meeting. Additionally, the announcement was posted in both video and text form on each lab course's learning management site, along with a link to the Qualtrics survey. Students were offered a nominal amount of extra credit for accessing each survey. In the pre-term surveys, students were randomly presented with ten MLLI items and asked to respond using the 0-100 scale. To provide additional response process data, students were also asked to describe why they selected each response value. For each of the 30 MLLI items, 6 interview responses and approximately 100 open ended written responses were collected.

Quantitative Data Collection

Quantitative data were collected from Portland State University (PSU) students taking a first term general chemistry laboratory course in the fall of 2021 or a first term organic laboratory course in the winter of 2022. Additionally, students taking a first semester general or organic chemistry laboratory course at East Carolina University (ECU) and students taking a first semester general chemistry laboratory course at San Diego State University (SDSU) in the fall of 2021 were also included in this portion of the study. All quantitative MLLI data were collected through the distribution of both a pre- and post-course survey via Qualtrics. For survey recruitment, a video announcement was pre-recoded by the first-author (E.B.V.) and presented by the graduate teaching assistants in each laboratory section. Additionally, the announcement was posted in both video and text form on each courses' learning management site, along with a link to the Qualtrics survey. PSU students were required to complete both the pre- and post-course surveys as graded course assignments, although allowing their data to be used for the proposes of this research project remained voluntary. Students at the other two institutions were offered a nominal amount of extra credit for accessing each survey. The pre-course survey was

distributed during the first two weeks of each laboratory course, while the post-course survey was distributed in the final two weeks of each laboratory course. To prevent item-order effects in the data, the MLLI items were randomized for each student. As suggested by the MLLI authors (Galloway and Bretz, 2015a), a check item was included in each survey to allow for the removal of student responses who were not reading the items. Before analyzing any quantitative data, responses were cleaned by removing any duplicate participant responses or responses from participants who incorrectly responded to the check item. A table of cleaned response totals for the quantitative data collection can be found in Table 1. Additionally, self-reported demographics data for the cleaned student responses are provided in Table S1 in the Supporting Information.

Table 1. Cleaned student response totals for data collected with the MLLI

Course	Pre-Term Responses (n)	Post-Term Responses (n)
Portland State University (PSU)		
General Chemistry	192	169
Organic Chemistry	109	96
East Carolina University (ECU)		
General Chemistry	285	167
Organic Chemistry	97	95
San Diego State University (SDSU)		
General Chemistry	186	95
Totals	869	622

Analysis Methods

Qualitative Data Analysis

The interview transcripts and written item responses were analyzed to determine if participants were interpreting and responding to MLLI items as intended. To do this, two researchers individually read through each participant's interview transcript or free responses and flagged items that did not appear to be functioning properly. Items were flagged if: 1) the participant's explanation did not match the selected numerical response, (i.e., if a participant selected a numerical value on the "Agree" side of the scale, then their explanation of why they chose that response should also indicate that they agree with the item), 2) the participant expressed confusion about the meaning of the item (e.g., if they asked for clarification or

indicated that they were unsure how to respond), 3) their explanation indicated that they interpreted the item differently from its intended meaning, 4) they responded that the item was not relevant to them personally or to the environment in which the question was posed, and/or 5) they interpreted items as being redundant. The two researchers then came together and discussed the clarity and relevance of any flagged items, in order to come to a consensus. Results from this analysis were used to provide qualitative support for the removal of poorly functioning items.

Quantitative Data Analysis

After cleaning, the remaining student responses from each of the three institutions were combined into an aggregated data set, which was then randomized and split into two approximately equal halves. The first of set of data (termed the 'training' data set) was analyzed through item descriptive statistics (Supporting Information, Table S2) and initial factor analysis, while the second set (termed the 'testing' data set) was used for cross-validation with the final model. All negatively worded items were reverse coded before analysis.

Confirmatory Factor Analysis

To provide evidence in support of the structural validity and single-administration reliability for the data collected with the MLLI, a variety of CFA models were evaluated. All CFAs were completed using the statistical program R (version 4.2.0 (2022-04-22)) with the package lavaan (version 0.6-11). Maximum likelihood with Satorra-Bentler adjustment and robust standard errors were used to account for any non-normality of the data (Satorra & Bentler, 1994). Fit statistics were calculated and interpreted for goodness of the data-model fit, where the guidelines for good fit are CFI & TLI ≥ 0.95, RMSEA ≤ 0.06, SRMR ≤ 0.08; and adequate fit are CFI & TLI ≥ 0.90, RMSEA ≤ 0.08, SRMR ≤ 0.10 (Hair et al., 2010; Hu & Bentler, 1999; Marsh et al., 2004; Schweizer, 2010; Brown and Cudeck, 1993; Kline, 2005).

Model Modification

Poor data-model fit, modification indices, and low item loadings were used to flag items that may have been functioning poorly in the training data set. These data were used in parallel with qualitative response process data to support the removal of items when necessary. Once items were removed, data from the reduced set of MLLI items was then reanalyzed via CFA to assess data-model fit. To provide evidence of cross validation for the final model and reduced set of items, the data-model fit of the testing data set was evaluated. Cross-validation is one way to

provide support for the data-model fit of modified models using 'unique' data (i.e., data that was not directly used in making the modifications) (Koul et. al., 2018).

McDonalds Omega

Using the testing data set, the single-administration reliability of each unidimensional factor (evaluated via CFA) was assessed using McDonald's Omega. Values for omega range from 0 to 1, with 1 indicating that all of the observed variance is from the latent construct. Therefore, a high omega value (> 0.7) provides evidence to support of the internal consistency of the items (McDonald, 1999).

Results and Discussion

Before the data collected in this study was used to investigate evidence in support of validity and reliability, the 869 cleaned student responses to the pre-course survey were aggregated, randomized, and split into two approximately equal halves. The first set of data (termed the 'training' data set) included 434 responses and was analyzed through item descriptive statistics and factor analysis, while the second set (termed the 'testing' data set) included the remaining 435 responses and was used for cross-validation. Additionally, because data collected with MLLI is theorized to have the same internal structure for both the pre-course and post-course administrations (Galloway and Bretz, 2015a), the post data set was used as an additional source of cross-validation.

A Priori Model Evaluation

In order to support the calculation of scores for each latent construct measured by the MLLI, the first step was to investigate the structure of the MLLI data from the sample of students in this study. Using the training dataset, the two structures previously suggested by the authors of the MLLI (Galloway and Bretz, 2015a) were tested via CFA to determine which structure was the most appropriate for use with our data. The first model tested (Model A) was the three-factor structure (Figure 1), where each of the 30 MLLI items was qualitatively sorted into cognitive, affective, and cognitive/affective groups using meaningful learning as a theoretical framework. As shown in Table 2, Model A showed evidence of poor data-model fit, as each of the fit indices fell outside of their recommended ranges. Additionally, the correlation between the affective and cognitive/affective factors in Model A was 0.949. Factor correlations

close to 1 indicate that these items may be measuring the same construct (Brown, 2015). The next model tested (Model B) was the two-factor structure identified by the MLLI authors via EFA (Figure 1), in which each factor contained only positively or negatively worded items. The two-factor model only included 28 items, as the items reported by Galloway and Bretz that did not cleanly load on a single factor were removed before analysis. The removal of these items is necessary, as items that are cross loading on multiple factors, or do not load on either factor, do not provide support for determining a 'unique' score for each latent variable (Li, et al., 2020). Data-model fit statistics (Table 2) for Model B were much closer to the acceptable ranges than for Model A, but all still fell outside the suggested cutoffs for acceptable data-model fit. Factor loadings for all items in Model A and Model B can be found in the Supporting Information (Tables S3 and S4). As neither Model A nor Model B resulted in acceptable data-model fit, the next step was to identify poorly functioning items and possible modifications to the MLLI models which may better represent the data.

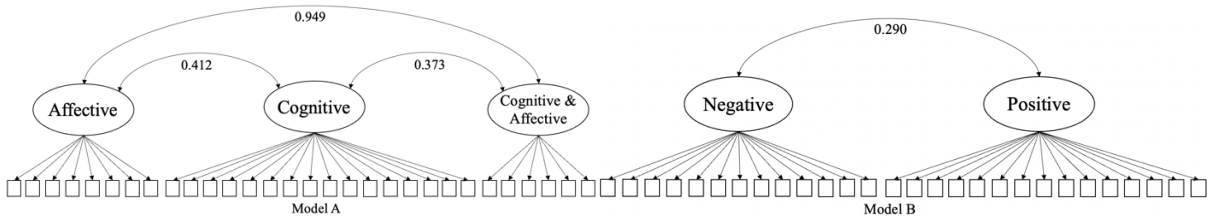

Figure 1. *A priori* structures of the MLLI data. Model A - three-factor structure suggested by Galloway and Bretz (2015a), where each of items was qualitatively grouped into cognitive, affective, and cognitive/affective categories using meaningful learning as a theoretical framework. Model B – two-factor structure of positively and negatively worded items suggested by Galloway and Bretz (2015a) based on EFA results.

Table 2. Data-model fit statistics for *a priori* structures for the MLLI (n = 434). Italic values indicate the results met the suggested cutoff criteria for adequate fit (CFI & TLI ≥ 0.90, RMSEA ≤ 0.08, SRMR ≤ 0.10). Bold values indicate that the results met the suggested cutoff criteria for good fit (CFI & TLI ≥ 0.95, RMSEA ≤ 0.06, SRMR ≤ 0.08) (Hair et al., 2010; Hu & Bentler, 1999; Marsh et al., 2004; Schweizer, 2010; Brown and Cudeck, 1993; Kline, 2005).

Model	χ^2 (df)	p-value	CFI	TLI	RMSEA [90% CI]	SRMR
Model A	2557.692 (402)	<0.001	0.534	0.496	0.120 [0.116-0.125]	0.188
Model B	958.900 (349)	<0.001	0.860	0.848	*0.069 [0.064-0.075]*	*0.081*

Model Modification and Analysis Process

Data analysis investigating the structure of the MLLI began by using confirmatory factor analysis to evaluate each of the two *a priori* models (e.g., Model A and Model B, Figure 1). Because evidence of poor data-model fit was found for each of the models, further analysis steps (outlined in Figure 2 and detailed in the subsequent sections) were deemed necessary. The next

step in this analysis was to identify poorly functioning items and possible improvements to the *a priori* models. To do so, qualitative response process validity data in the form of cognitive interviews and free response items were analyzed and used to support decisions and further analyses. Items identified to not be functioning properly through qualitative analysis were then removed, and the data-model fit for each of the reduced factor structures was investigated via CFA. Additionally, a third possible factor structure was also tested. Once the best fitting model among these three possible factor structures was identified, further improvements to the MLLI data structure were then assessed through the investigation of modification indices. Suggested modifications were further investigated using the qualitative response process evidence. The factor structure of the final reduced set of MLLI items was then investigated through CFA and cross-validated using the testing-set and post-course data.

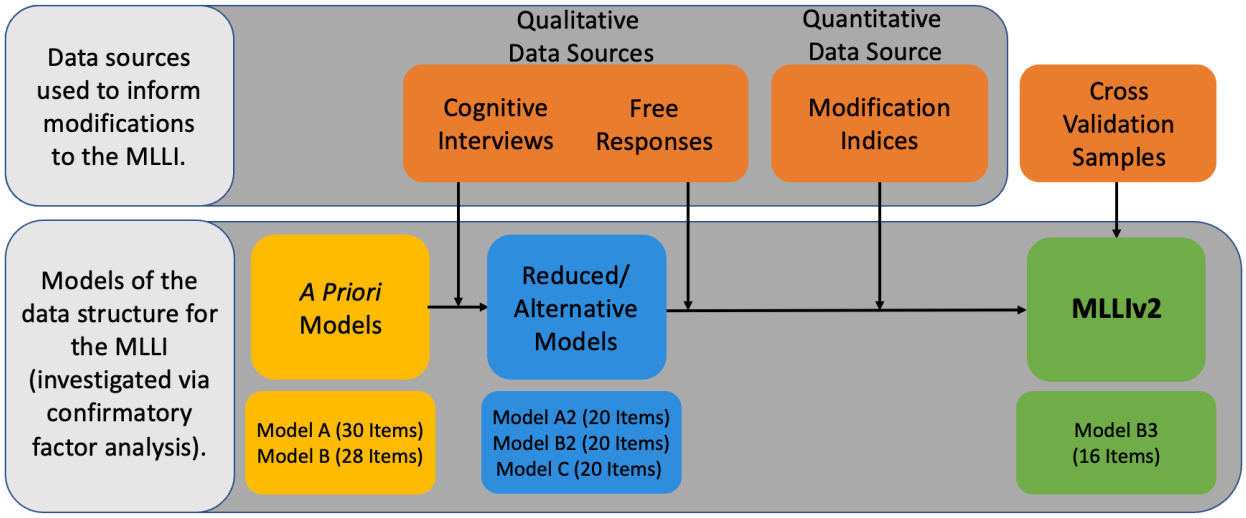

Figure 2. Flow chart of analysis steps to taken to evaluate and modify the MLLI data structure for the population of interest in this study.

Initial Item Removal

Because evidence of poor data-model fit was found for each of the *a priori* models (i.e., Model A and Model B, Figure 1), response process data were used to explore which items may have been functioning poorly. In total, ten items were removed based on response process validity evidence. All removed items, as well as their reason for removal and a representative student response can be found in Table 3. The reasons for item removal included students not understanding the intent/context of the item, students finding items to be ambiguous and/or

double barreled, students believing the negatively worded items represented positive aspects of participating in a laboratory course, and inconsistent response scale usage. For example, the item "…to be nervous when handling chemicals" was removed because many students believed that the item represented a positive aspect of their laboratory experience, while the authors of the item had categorized it as negative. One student wrote, "It's good to be nervous in my perspective, provides a sense of respect for the chemicals, and safety for yourself and people around you in a laboratory setting." Alternatively, the item "…to focus on procedures, not concepts" was removed because students demonstrated inconsistent use of the response scale, which may have been due to the double-barreled nature of the item. For example, one student responded to the item with a numerical value of 0 and stated that "Both [procedures and concepts] are equally important," while their classmate responded with a numerical value of 50, but provided a similar explanation: "I want to fully understand both. 50/50." While some variation in student responses is expected, students selecting numerical values on drastically different parts of the 0-100 response scale, while reporting nearly identical reasoning, indicates that this item is not functioning properly. Of the ten items removed at this stage, six were designated as cognitive, two were designated as affective, and two were designated as cognitive/affective. Four of the items were positively worded and six of the items were negatively worded.

Table 3. MLLI items removed using response process validity and reasons for item removal. Assigned categories for items include Cognitive (C), Affective (A), and Cognitive/Affective (C/A). Item wording includes positively worded items (+) and negatively worded items (-).

Reason for Removal	Item (C, A, C/A) (+, -)	Response Value	Representative Quote(s)	Discussion
Written as negative, considered positive by students	to worry about getting good data. (C/A) (-)	100	I don't know if worry is the right term, but I will be focused on getting good as well as correct data for the experiments.	Students discussed that they did not worry about getting good data in a negative sense, but that they did believe that collecting good or correct data was an important part of being successful in their laboratory course.
	to be nervous when handling chemicals. (A) (-)	100	It's good to be nervous in my perspective, provides a sense of respect for the chemicals, and safety for yourself and people around you in a laboratory setting.	Student found this item similar to "to worry about getting good data." In the case of this item, students once again reported that collecting good or correct data was an important part of being

				successful in their laboratory course.
	to worry about the quality of my data. (C/A) (-)	100	Quality of data is just as important as any other variable in an experiment so I will be sure to make sure it is good.	Students believed that being somewhat nervous when working with chemical was a positive thing, as it meant that they were being appropriately cautious and safe around potentially harmful materials.
Ambiguous	the procedures to be simple to do. (C) (-)	100	100% because while the procedure may be difficult there should be some form of teaching to make it seem simple.	Students had a difficult time interpreting the word simple in this item. Most students believed that procedures should be doable/ completable, but it was unclear whether students appropriately understood the language of this item.
		53	Yes and no. Sometimes things are a bit confusing but I get through them. Not all projects are going to be simple.	
	to think about chemistry I already know. (C) (+)	50	Not very often, as I don't always feel what we do in lab pertains to class.	Many students had a difficult time interpreting this item, which was intended to probe if students were using their existing knowledge of chemistry in their laboratory course. Instead, students focused on the fact that they hoped to learn new chemistry content, and that the lecture course content may not always align with the laboratory course content.
		11	No, I expect to learn chemistry I don't know.	
	to worry about finishing on time. (A) (-)	0	I do not expect that at all. We usually have the full week to submit the report.	Students stated that they were not concerned about laboratory timelines or due dates. This item was also ambiguous for students; some participants referred to in laboratory activities, while others discussed the time allotted outside of the laboratory to complete lab reports.
		7	I feel like 3 hours is enough time to finish a lab.	
Double Barreled/ Inconsistent Use of Response Scale	to "get stuck" but keep trying. (C) (+)	39	If I do get stuck I know I'll keep going, but I hope to not actually get stuck.	Students struggled with the two-part nature of this item, often considering them separately when describing their response reasoning. Additionally, students with similar reasonings selected numerical values in very different parts of the scale, highlighting students' inconsistent scale usage for this item.
		5	I don't anticipate getting stuck but if I do obviously I will keep at it until I have a firm understanding of the experiment.	
	to focus on procedures, not concepts.	0	Both [procedures and concepts] are equally important.	Many students described that it was important to focus on both procedures and concepts in their

	(C) (-)			laboratory courses. That said, scale use was inconsistent, where some students with this reasoning chose to disagree with the item (0), while others chose neither agree nor disagree (50).
		50	I want to fully understand both. 50/50	
	to make mistakes and try again. (C) (+)	22	I hope I will not make too many mistakes but if I do I will work to correct them.	Students found this item similar to ‘to "get stuck" but keep trying.’ They once again struggled with the two-part nature of this item. As seen in the student quotes, students with similar reasonings selected numerical values in very different parts of the scale, highlighting students’ inconsistent scale usage for this item.
		49	I expect to make some mistakes so I can learn from them, but I don't want to make so many that I end up frustrated in the end.	
		100	I always make mistakes but I always try to mediate them. I do not like getting things wrong, so I like to try to prevent mistakes in the future.	
Incorrect Interpretation of Item	to consider if my data makes sense. (C) (+)	42	I will always ask to make sure I’m doing things correctly.	While this item was intended to probe if students thought critically about their data in order to consider if it was reasonable, students frequently interpreted the ‘makes sense’ wording in this item as ‘correct’. Students highlighted a drive for ‘accurate’ data without an understanding of the concepts behind it.
		70	It's important to double-check everything to make sure everything is correct.	
		100	I always double-check to see if my data makes any sense even when I don't understand what I am doing.	

Evaluation of A Priori Structures with Reduced Itemset

After removing items using response process validity evidence, two reduced models (Figure 3) were reassessed using CFA. The first reduced model investigated was the 20-item cognitive, affective, and cognitive/affective three-factor model (Model A2). Although the data-model fit statistics were nominally improved from the 30-item Model A, they still fell far below acceptable data-model fit (Table 4). Of additional concern, the correlation between each of the three factors ranged between 0.900 and 1.171, once again indicating that these items may be measuring the same construct. The correlations of 1.171 between the affective and cognitive/affective factors and 1.015 between the cognitive and cognitive/affective factors are especially concerning, as they are over 1.0. Often, correlations between latent factors that are greater than 1.0 indicate a misspecification in the model (Dillon, et al., 1987). The 20-item two-

factor, positive/negative, reduced model (Model B2) was then assessed. The data-model fit statistics showed evidence of improved fit over Model B, and all of the fit statistics fell within the ranges for adequate data-model fit. (Table 4). Factor loadings for all items in Model A2 and Model B2 can be found in the Supporting Information (Tables S3 and S4).

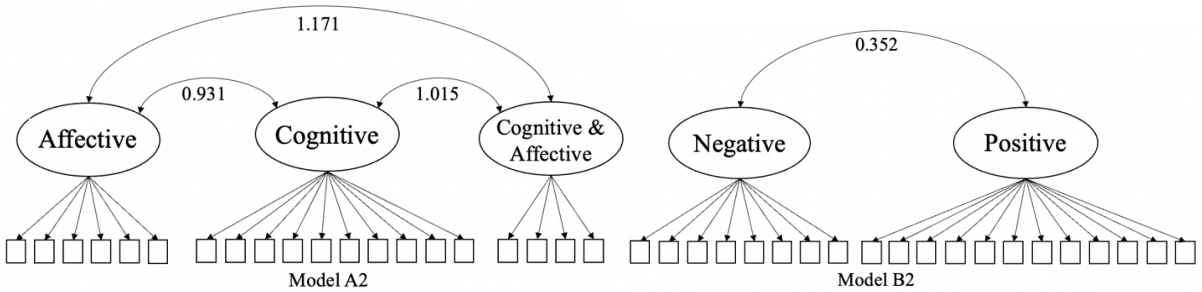

Figure 3. Twenty item reduced structures for the MLLI data, item removal based on analysis of response process validity data. Model A2 - three-factor structure with items grouped into cognitive, affective, and cognitive/affective categories. Model B2 - two-factor structure with items grouped by positively-worded and negatively-worded items.

In addition to the three- and two-factor correlated structures published by the original authors (i.e., Models A and B), there is a possible alternative structure which may represent the intended cognitive, affective, and cognitive/affective groupings, while also accounting for item polarity. Including positively and negatively worded items in a survey, as Galloway and Bretz have done in the MLLI, can encourage participants to read items more carefully and think about their responses, instead of simply responding agree to all items (Zeng, et al., 2020). While Galloway and Bretz interpreted the two-factor structure as "items contributing to meaningful learning" (positively worded items) and "items inhibiting meaningful learning" (negatively worded items) (Galloway and Bretz, 2015a), it is possible that the difference in how students respond to positively versus negatively worded items is not due to meaningful learning differences, but to differences in how students utilized the response scale for these types of items. (Zeng, et al., 2020). For example, a participant may be more likely to select "strongly agree" when responding to a positively worded item, but be less likely to "strongly disagree" with a similarly worded negative item, even if the items are related to the same latent construct. Therefore, given the potential for response bias, adding a negative method factor to the originally proposed three-factor structure could simultaneously account for the negative item grouping, while maintaining the three originally intended groupings for the MLLI. When the alternative model (Model C, Figure 4), with a negative method factor added to the original item groupings, was evaluated, the data-model fit (Table 4) was found to improve from that of the reduced three-

factor model (Model A2). Factor loadings for all items in Model C can be found in the Supporting Information (Table S5). While improved, the data-model fit statistics for Model C still fell outside the suggested acceptable range, leaving the reduced two-factor model (Model B2) as the best option for exploring further model modification.

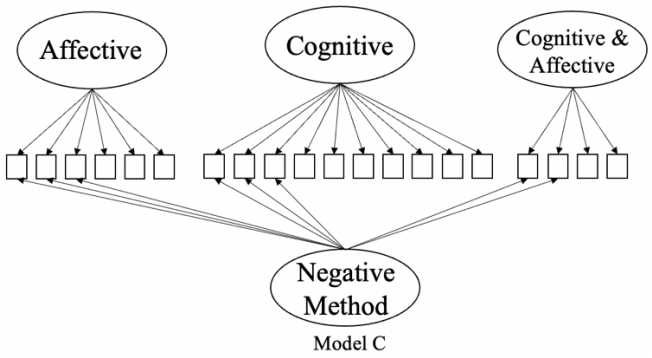

Figure 4. Twenty item alternative structure for the MLLI data. Model C - three-factor structure with items grouped into cognitive, affective, and cognitive/affective categories, with an added negative method factor.

Table 4. Data-model fit statistics for the reduced 20 item MLLI (n = 434). Italic values indicate the results met the suggested cutoff criteria for adequate fit (CFI & TLI ≥ 0.90, RMSEA ≤ 0.08, SRMR ≤ 0.10). Bold values indicate that the results met the suggested cutoff criteria for good fit (CFI & TLI ≥ 0.95, RMSEA ≤ 0.06, SRMR ≤ 0.08) (Hair et al., 2010; Hu & Bentler, 1999; Marsh et al., 2004; Schweizer, 2010; Brown and Cudeck, 1993; Kline, 2005).

Model	χ^2 (dof)	p-value	CFI	TLI	RMSEA [90% CI]	SRMR
Model A2	1499.936 (167)	<0.001	0.585	0.527	0.150 [0.143-0.157]	0.159
Model B2	403.867 (169)	<0.001	*0.924*	*0.915*	*0.064 [0.056-0.072]*	**0.063**
Model C	909.618 (162)	<0.001	0.760	0.719	0.116 [0.108 -0.123]	0.215

Further Model Modifications

Once the best fitting model was identified (i.e., Model B2), its modification indices were investigated. Modification indices can be examined to determine if there are suggested model alterations that may be appropriate, such as the correlation of error terms or the association of an item with a different factor (Knekta et al., 2019). When supportive qualitative evidence exists, these suggested modifications can be made to the model to improve the data-model fit. Analysis of Model B2 revealed five pairs of items with large modification indices for the correlation of error terms. Correlated errors frequently exist between items with similar wording or content (Knekta et al., 2019). Upon further investigation of the response process validity data for these items, it was found that students were interpreting four of these pairs of items similarly (Table 5). In each of these cases, qualitative response process data were used to determine which of the two items should be removed. If it was not obvious which item should be removed, the item with the

higher factor loading was retained. For example, analysis of modification indices revealed that the items, "…to learn critical thinking skills" and "…to learn problem solving skills" had large modification indices suggesting the correlation of error variances. Returning to the response process validity data for these items, it was found that students could not easily distinguish the difference between 'critical thinking' and 'problem solving'. For instance, one student stated that "I strongly believe that one of the most important skills you can learn [is] to critically think, and problem solving goes hand in hand with that." Additional discussion of the reasons for removal for the remaining items can be found in Table 5.

Table 5. MLLI items flagged using modification indices and discussion of the similarities between items. Assigned categories for items include Cognitive (C), Affective (A), and Cognitive/Affective (C/A). Item wording includes positively worded items (+) and negatively worded items (-).

Item Pairs in Which One Item Was Removed		
Removed Item **(C, A, C/A) (+, -)**	**Retained Item** **(C, A, C/A) (+, -)**	**Discussion of Items/ Reason for Removal**
to learn critical thinking skills. (C) (+)	to learn problem solving skills. (C) (+)	Students could not easily distinguish the difference between 'critical thinking' and 'problem solving'.
to be nervous about making mistakes. (A) (-)	to feel intimidated. (A) (-)	Students who agreed with these items typically discussed feeling overwhelmed by new chemistry content/material. Students who disagreed with these items discussed difficulties in lab such as making mistakes being a part of the learning process.
to think about what the molecules are doing. (C) (+)	to use my observations to understand the behavior of atoms and molecules. (C) (+)	Students identified that these two items are similarly worded. For each of these items, students discussed that understanding what the molecules are doing is an important part of laboratory activities.
to be confident when using equipment. (A) (+)	to develop confidence in the laboratory. (A) (+)	Students identified that these two items are similarly worded. For each of these items, students discussed a desire to become more confident in their laboratory courses.
Item Pairs in Which Both Items Were Retained		
Retained Item **(C, A, C/A) (+, -)**	**Retained Item** **(C, A, C/A) (+, -)**	**Discussion of Items/ Reason for Retention**
to learn chemistry that will be useful in my life. (C/A) (+)	to be excited to do chemistry. (A) (+)	While each of these items have an affective component, students' qualitative responses to these items did not did not reveal substantial overlap in their interpretation of the items.

After removing 4 items using a combination of modification indices and response process validity evidence, the final 16-item two-factor model, which included seven items on the negative factor and nine items on the positive factor (Model B3, Figure 5), was reanalyzed using CFA. Factor loadings for all items in Model B3 can be found in the Supporting Information (Table S6). Fit statistics for this model fell within the acceptable to good data-model fit ranges (Table 6) and provided structural validity support for this two-factor model. Furthermore, this evidence provides support for the reporting of positive and negative scale scores using the 16 remaining MLLI items.

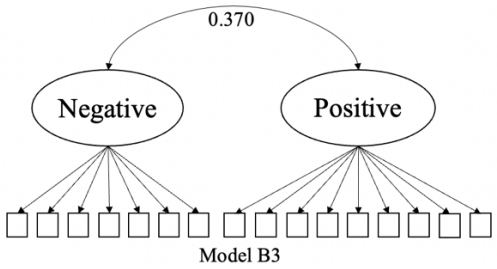

Figure 5. Two-factor structure for the final 16 item MLLI with positive and negative factors.

Table 6. Data-model fit statistics and for the final 16 item MLLI (Model B3) (n = 434). Italic values indicate the results met the suggested cutoff criteria for adequate fit (CFI & TLI ≥ 0.90, RMSEA ≤ 0.08, SRMR ≤ 0.10). Bold values indicate that the results met the suggested cutoff criteria for good fit (CFI & TLI ≥ 0.95, RMSEA ≤ 0.06, SRMR ≤ 0.08) (Hair et al., 2010; Hu & Bentler, 1999; Marsh et al., 2004; Schweizer, 2010; Brown and Cudeck, 1993; Kline, 2005).

χ^2(df)	p-value	CFI	TLI	RMSEA [90% CI]	SRMR
229.454 (103)	<0.001	*0.946*	*0.938*	**0.059 [0.049-0.069]**	**0.056**

Reliability Evaluation

To provide single-administration reliability support for the individual MLLI factors from Model B3 (Figure 5), single factor CFAs were conducted to evaluate the unidimensionality of each factor. Factor loadings for the individual positive and negative factors can be found in the Supporting Information (Tables S7). Evidence of good data-model fit was found for each factor (Table 7). As each factor was found to be unidimensional, an investigation of internal reliability using McDonalds omega was conducted and acceptable values of single administration reliability (>0.7) were found.

Table 7. Data-model fit statistics for the two individual positive and negative factors that make up the final 16 item MLLI (Model B3) (n = 434). Italic values indicate the results met the suggested cutoff criteria for adequate fit (CFI & TLI ≥ 0.90, RMSEA ≤ 0.08, SRMR ≤ 0.10). Bold values indicate that the results met the suggested cutoff criteria

for good fit (CFI & TLI ≥ 0.95, RMSEA ≤ 0.06, SRMR ≤ 0.08) (Hair et al., 2010; Hu & Bentler, 1999; Marsh et al., 2004; Schweizer, 2010; Brown and Cudeck, 1993; Kline, 2005).

Single Factor	χ^2(df)	p-value	CFI	TLI	RMSEA [90% CI]	SRMR	Omega
Positive	63.816 (36)	<0.001	**0.959**	*0.946*	*0.066 [0.045-0.087]*	**0.039**	0.86
Negative	37.808 (14)	0.001	**0.979**	**0.969**	*0.072 [0.045-0.100]*	**0.030**	0.90

Cross-validation of Two-Factor Model

Any time a model is modified, cross-validation is suggested in order to provide further supporting evidence for the structural validity of the modified measure (Kline, 2011). In this study, evidence of cross-validation was provided by assessing the data-model fit of both the testing data set and the data collected with the post-course assessment (Table 8). The testing data set maintained evidence of acceptable to good data-model fit and the post-term data provided evidence of acceptable data-model fit. Further evidence of cross-validation was provided through an additional data collection using the 16-item MLLIv2 in the 2022/2023 academic year (Table 8). This data collection took place in general and organic chemistry laboratory courses at Portland State University and utilized the same data collection process described in the Methods: Participants and Data Collection portion of this manuscript. The data collected with the 16-item MLLIv2 once again exhibited evidence of acceptable to good data-model fit.

Table 8. Data-model fit statistics for the final 16 item MLLI (Model B3, two factor model). Italic values indicate the results met the suggested cutoff criteria for adequate fit (CFI & TLI ≥ 0.90, RMSEA ≤ 0.08, SRMR ≤ 0.10). Bold values indicate that the results met the suggested cutoff criteria for good fit (CFI & TLI ≥ 0.95, RMSEA ≤ 0.06, SRMR ≤ 0.08) (Hair et al., 2010; Hu & Bentler, 1999; Marsh et al., 2004; Schweizer, 2010; Brown and Cudeck, 1993; Kline, 2005).

Data Set	n	χ^2 (df)	p-value	CFI	TLI	RMSEA [90% CI]	SRMR
[a]Training	434	229.454 (103)	<0.001	*0.946*	*0.938*	**0.059 [0.049-0.069]**	**0.056**
Testing	435	248.979 (103)	<0.001	*0.941*	*0.932*	*0.062 [0.052-0.072]*	**0.062**
Post	622	412.073 (103)	<0.001	*0.899*	0.883	*0.069 [0.063-0.076]*	**0.063**
Pre-MLLIv2	395	311.725 (103)	<0.001	*0.906*	0.891	*0.079 [0.069-0.089]*	**0.068**
Post-MLLIv2	340	242.218 (103)	<0.001	*0.931*	*0.920*	*0.069 [0.058-0.081]*	**0.060**

[a]Training set data were initially reported in Table 4 and are displayed here for comparison purposes.

Conceptualization of Constructs Measured by the MLLIv2

The data collected with the MLLIv2 provides evidence for a two-factor model. The authors hypothesize that these factors likely represent student expectations that contribute to meaningful learning and student expectations that detract from meaningful learning. While these

two opposing item categories were originally described by the authors of the MLLI after conducting an exploratory factor analysis (Galloway and Bretz, 2015a), the factor structure of the MLLIv2 can be contextualized through Ausubel and Novak's theories of education.

Ausubel's assimilation theory suggests that cognitive learning exists on a continuum from rote learning to meaningful learning, where rote learning is described as "arbitrary, verbatim incorporation of new information into cognitive structure" and meaningful learning is described as when "new knowledge is consciously linked to existing specifically relevant concepts and propositions in cognitive structure and incorporated into these concepts" (Novak, 1980; Ausubel et al., 1978). Novak's theory of education builds on Ausubel's by suggesting that meaningful learning is dependent not only on the cognitive domain, but instead "underlies the constructive integration of thinking (cognitive domain), feeling (affective domain), and acting (psychomotor domain) leading to empowerment for commitment and responsibility" (Novak, 2010). Ausubel and Novak agree that in order for meaningful learning to occur (as opposed to rote learning), the learner "must actively try to link new knowledge with existing, relevant knowledge" (Novak, 1980; Ausubel et al., 1978). When examining the items retained in the MLLIv2, it is possible that the positively worded items represent student expectations that support the occurrence of meaningful learning. For example, items such as "…[I expect] to learn chemistry that will be useful in my life", "…[I expect] to experience moments of insight", and "…[I expect] to interpret my data beyond only doing calculations" all directly prompt students to consider the relationship between their existing life experience/knowledge and their learning experiences within the chemistry laboratory course. While the negatively worded items included in the MLLIv2 may not directly reflect rote learning, they may represent student expectations that detract from the occurrence of meaningful learning. For example, items such as "…[I expect] to feel unsure about the purpose of the procedures", "…[I expect] to be confused about the underlying concepts", and "…[I expect] to feel frustrated" may prompt students to reflect on possible difficulties (e.g., uncertainty, confusion, frustration) that could inhibit the integration of new knowledge into existing knowledge structures.

Conclusions

The aim of this study was to investigate the validity and reliability evidence related to data collected with the Meaningful Learning in the Laboratory Instrument (MLLI) prior to the

use of its data as part of a larger study. Supportive evidence for the structure of data provides greater meaning for the reporting and interpretation of scores from the latent constructs measured by the instrument. To address research question one, *What evidence of validity and reliability supports interpreting data collected with the MLLI as measures of student expectations in lower division undergraduate laboratory courses?*, data collected with the instrument was investigated for evidence in support of structural validity using two *a priori* models for the factor structure of the instrument (Galloway and Bretz, 2015a). This analysis found evidence of poor data-model fit for each of the two tested models (Model A and Model B). Because supportive evidence of structural validity and was not found for the data collected with the MLLI, research questions two and three were also addressed.

Research question two, *If insufficient evidence is found, what modifications could potentially improve the MLLI data with this population?*, was addressed next. Cognitive interviews and free response items, in addition to data from the initial CFAs, were used to investigate and modify the two *a priori* models through the removal of 10 poorly-functioning items. Additionally, a third alternative factor structure, which included a negative method factor, was investigated. Results from these analyses indicated that the 20 item two-factor model with positively and negatively worded item groupings (Model B2) showed evidence of adequate to good data-model fit, while the other two modified models (Model A2 and C) showed poor data-model fit. Therefore, Model B2 was further explored through evaluating modification indices. This analysis revealed four pairs of items with modification indices that suggested the correlation of error variances. In each case, qualitative response process data were used to provide evidence for the removal of one of the two paired items, resulting in a final set of 16 items.

Because modifications were made to the MLLI data structure, research question three was also addressed: *If modifications are necessary, what evidence of validity and reliability supports interpreting data collected with the modified instrument?* Results from CFA using the final 16 item two-factor model (Model B3) showed evidence of adequate to good data-model fit. Additionally, evidence of both internal structure validity and single administration reliability was found for each of the final two individual factors. Lastly, cross-validation in the form of results from CFA using the testing-set data and post-course data showed evidence of adequate to good data-model fit. The 16 items, deemed the MLLIv2, can be found in Table 9. The structure of the data from these items leads to scale scores that represent student expectations which may

contribute to (positive) and detract from (negative) meaningful learning. These two opposing item categories were originally described by the authors of the MLLI after conducting an exploratory factor analysis (Galloway and Bretz, 2015a). Additionally, this factor structure can be contextualized through Ausubel and Novak's theories of education (Novak, 1980; Ausubel et al., 1978).

While the theoretical work provided by Ausubel and Novak may support users of the MLLIv2 in the interpretation of these constructs, additional studies related to test content would be useful to provide more substantial evidence for the identities of the constructs measured by the MLLIv2. Ongoing studies from the larger project which this study is part of, are investigating the relations between MLLIv2 scores and other theoretically related variables. Therefore, future manuscripts from this project will provide additional validity evidence, in the form of relations to other variables, to support the interpretation of MLLIv2 scores.

Table 9. MLLIv2 factors and items

Meaningful Learning in the Laboratory Instrument – Version 2 (MLLIv2)	
Pre-Course Item Stem	*When performing experiments in my chemistry laboratory, I expect…*
Positive Items	
to learn chemistry that will be useful in my life.	
to make decisions about what data to collect.	
to experience moments of insight.	
to be excited to do chemistry.	
to develop confidence in the laboratory.	
to interpret my data beyond only doing calculations.	
to use my observations to understand the behavior of atoms and molecules.	
to be intrigued by the instruments.	
to learn problem solving skills.	
Negative Items	
to feel unsure about the purpose of the procedures.	
to be confused about how the instruments work.	
to feel disorganized.	
to be confused about the underlying concepts.	
to be frustrated.	
to feel intimidated.	
to be confused about what my data mean.	

Limitations

The response process validity data included in this study was collected at a single institution. Thus, the qualitative results may not be generalizable to other student populations. Additionally, the qualitative portion of the study was completed in virtual laboratory courses (during the Covid-19 shutdown), while the quantitative data collection was conducted during in-person courses. The authors acknowledge that students' responses to open ended items and interview questions related the MLLI may have been influenced by the virtual learning format. For this reason, extra care was taken in the cognitive interviews (in the form of follow-up questions) to ensure that students were appropriately interpreting and responding to the MLLI items.

Implications for Research

The goal of this project was to investigate evidence of validity and reliability for data collected with the MLLI, a widely used instrument in the field of chemistry education. This study found validity and reliability evidence that supported a 16 item two-factor structure, with positively and negatively worded item groupings, deemed the MLLIv2. This evidence provides psychometric support for the reporting of positive and negative factor scores from the MLLIv2 for data collected with our population. As many prior studies with the MLLI have reported scores for the three proposed item categories (i.e., cognitive, affective and cognitive/affective), this study evaluated several three-factor models but did not find sufficient evidence to support the intended theoretical structure. While evidence of the proposed three item category structure was not found for data collected in the environment in which this study was conducted, that does not mean that evidence for this structure did not exist in previous studies. That said, caution should be used when interpreting the results of studies where the data structure is not supported.

The results of this study highlight an important aspect of education research. Before a scale score can be calculated and meaningfully interpreted, evidence to support the structural validity and reliability should first be assessed. This need exists, both when an instrument is being developed and when it is being used in a new environment. As described in the *Journal of Chemical Education*, "Chemical Education Research (CER) has come a long way as a research discipline over the past century, moving away from personal empiricism (i.e., sole reliance on one's personal experience to provide advice and recommendations) to empirical investigations (i.e., systematic collection of valid and reliable evidence informed by theoretical

perspectives). Reviewers systematically requested evidence for the validity and reliability of the data collected, whether the data were collected with a new or already existing and published instrument" (Stains, 2022). More specific to validity, a recent editorial published in *Chemistry Education Research and Practice,* asserted that "Ultimately, the quality and transparency of evidence for validity plays a central role in evaluating the appropriate uses of the data collected. Researchers that analyze quantitative data are therefore encouraged to incorporate an explicit account on the evidence for validity of the data collected" (Lewis, 2022). Together, these assertations by two of the leading publications in the field of chemistry education research highlight the necessity of assessing evidence of validity and reliability when developing or using an instrument in a new environment.

The MLLIv2 described in this article could be used by future researchers in a variety of ways. If sufficient evidence of measurement invariance is found (Rocabado et al., 2020), group comparisons of students' expectations which contribute to and/or detract from meaningful learning in their laboratory courses could be made. Comparisons between different course types (e.g., general vs. organic chemistry), course delivery methods (e.g., cookbook vs. inquiry style labs), and institutions could be of interest to education researchers and practitioners alike. For example, existing literature using the original 30 item version of the MLLI has suggested that organic chemistry students tend to have lower expectations scores (as measured by the MLLI) than general chemistry students, a trend that may be due to unmet laboratory expectations in previous STEM laboratory courses (Galloway & Bretz, 2015b). Revisiting this claim using the MLLIv2 could be an interesting avenue of investigation for chemistry education researchers.

Implications for Teaching

As students' expectations are theorized to impact their learning experience (DeKorver & Towns, 2015; Galloway and Bretz, 2015a; Wang et al., 2021), collecting feedback related to students' expectations, especially those related to meaningful learning, could provide instructors better insight into the perspectives of their students. Because the two factors of the MLLIv2 are proposed to represent student expectations that may encourage and detract from meaningful learning, data collected with the MLLIv2 may be used by practitioners to make informed adjustments to their laboratory activities, which may increase the likelihood of students' learning meaningfully in their chemistry laboratory courses.

The results of this study also highlight the importance of collaborations among researchers and practitioners. Practitioners who want to use the MLLIv2 to inform their teaching may want to work with a researcher who is interested in investigating the validity and reliability of the data collected with the instrument. Additionally, as used by George-Williams (2019), either the MLLI or the MLLIv2 may be useful for practitioners at the item level, by looking at how students' perspectives on a single item may change over time.

References

1. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, (2014), *Standards for Educational and Psychological Testing*, American Educational Research Association, Washington, DC.
2. Arjoon, J. A., Xu, X., Lewis, J. E., (2013), Understanding the State of the Art for Measurement in Chemistry Education Research: Examining the Psychometric Evidence. *Journal of Chemical Education,* **90**, 5, 536-545.
3. Ausubel D. P., (1962), A subsumption theory of meaningful verbal learning and retention. *Journal of Gen Psychology,* **66**, 213–224.
4. Ausubel D. P., (1963), The Psychology of Meaningful Verbal Learning, New York: Grune and Stratton.
5. Ausubel D. P., (1968), Educational Psychology: A Cognitive View, New York: Holt, Rinehart, and Winston.
6. Ausubel, D., Novak, J., Hanesian, H., (1978), Educational Psychology: A Cognitive View (2nd ed.). New York: Holt, Rinehart & Winston.
7. Barbera, J., Naibert, N., Komperda, R., Pentcost, T.C., (2021), Clarity on Cronbach's Alpha Use. *Journal of Chemical Education,* **98**, 2, 257–258.
8. Bretz, S. L., (2019), Evidence for the Importance of Laboratory Courses, *Journal of Chemical Education*, **96**, 2, 193–195.
9. Brown, T. A., (2015), Confirmatory factor analysis for applied research, New York: The Guilford Press.
10. Brown, M. W., Cudeck, R., (1993), Alternative ways of assessing model fit. In K. A. Bollen, J. S. Long (Eds.), Testing structural equation models (pp. 136–162). Newbury Park, CA: Sage.
11. Collins, D., (2003), Pretesting survey instruments: An overview of cognitive methods, *Quality of Life Research,* **12**, 229–238.
12. DeKorver, B. K., Towns, M. H., (2015), General Chemistry Students' Goals for Chemistry Laboratory Coursework, *Journal of Chemical Education,* **92**, 12, 2031–2037.
13. Deng, J. M., Streja, N., Flynn, A. B., (2021), Response Process Validity Evidence in Chemistry Education Research, *Journal of Chemical Education,* **98**, 12, 3656–3666.
14. Dillon, W. R., Kumar, A., Mulani, N., (1987), Offending Estimates in Covariance Structure Analysis: Comments on the Causes of and Solutions to Heywood Cases, *Psychological Bulletin,* **101,** 1, 126-135.

15. Galloway, K. R., Bretz, S., (2015a), Development of an Assessment Tool To Measure Students' Meaningful Learning in the Undergraduate Chemistry Laboratory, *Journal of Chemical Education*, **92**, 1149–1158.
16. Galloway, K. R., Bretz, S., (2015b), Measuring Meaningful Learning in the Undergraduate General Chemistry and Organic Chemistry Laboratories: A Longitudinal Study, *Journal of Chemical Education,* **92,** 12, 2019–2030.
17. Galloway, K. R., Bretz, S. L., (2015c), Using cluster analysis to characterize meaningful learning in a first-year university chemistry laboratory course, *Chemistry Education Research and Practice*, **16,** 879-892.
18. George-Williams, S. R., Karis, D., Ziebell, A. L., Kitson, R. R. A., Coppo, P., Schmid, S., Thompson, C. D., Overton, T. L., (2019), Investigating student and staff perceptions of students' experiences in teaching laboratories through the lens of meaningful learning, *Chemistry Education Research and Practice,* **20**, 187-196.
19. Grushow, A., Hunnicutt, S., Muñiz, M., Reisner, B. A., Schaertel, S., Whitnell, R., (2021), Call for Papers: Special Issue on New Visions for Teaching Chemistry Laboratory, *Journal of Chemical Education*, **98,** 3409-3411.
20. Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., (2010), Multivariate Data Analysis. 7th Edition, Pearson, New York.
21. Hu, L., Bentler, P. M., (1999), Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives, *Structural Equation Modeling*, **6,** 1-55.
22. Kline, R. B., (2005), Principles and practices of structural equation modeling. New York: Guilford Press.
23. Kline R. B., (2011), Principles and practice of structural equation modeling, 3rd Edition, New York: Guilford Press.
24. Knekta, E., Runyon, C., Eddy, S., (2019), One Size Doesn't Fit All: Using Factor Analysis to Gather Validity Evidence When Using Surveys in Your Research, *CBE Life Science Education,***18,** 1.
25. Komperda, R., Pentecost, T. C., Barbera, J., (2018), Moving beyond Alpha: A Primer on Alternative Sources of Single-Administration Reliability Evidence for Quantitative Chemistry Education Research, *Journal of Chemical Education,* **95,** 9, 1477-1491.
Koul, A., Becchio, C., Andrea, C., (2018), Cross-Validation Approaches for Replicability in Psychology, *Frontiers in Psychology,* **9,** 1117-1117.
26. Lau, P. N., Teow, Y., Low, X. T. T, Tan, S. T. B., (2023), Integrating chemistry laboratory–tutorial timetabling with instructional design and the impact on learner perceptions and outcomes, *Chemistry Education Research and Practice,* **24,** 12-35.
27. Lester, D., (2013), A Review of the Student Engagement Literature. *Focus on Colleges, Universities, and Schools,* **7,** 1.
28. Lewis, S. E., (2022), Considerations on validity for studies using quantitative data in chemistry education research and practice, *Chemistry Education Research and Practice,* **23,** 764-767.
29. Li, Y., Wen, Z., Hau, K. T., Yaun, K. H., Peng, Y., (2020), Effects of Cross-loadings on Determining the Number of Factors to Retain, *Structural Equation Modeling: A Multidisciplinary Journal,* **6,** 841-863.
30. Marsh, H. W., Hau, K. T., & Wen, Z. (2004), In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in

overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, **11**,3, 320-341.
31. McDonald, R. P., (1999), Test theory: A unified treatment. Mahwah, NJ: L. Erlbaum Associates.
32. Novak, J. D., (1980), Learning Theory Applied to the Biology Classroom, *The American Biology Teacher,* **42,** 5, 280-285.
33. Novak, J. D., (2003), The Promise of New Ideas and New Technology for Improving Teaching and Learning, *Cell Biology Education,* **2,** 122–132.
34. Novak, J. D., (2010), Learning, Creating, and Using Knowledge: Concept maps as facilitative tools in schools and corporations, *Journal of e-Learning and Knowledge Society,* **6,** 3, 21-30.
35. Rocabado, G. A., Komperda, R., Lewis, J. E., Barbera, J., (2020), Addressing Diversity and Inclusion through Group Comparisons: A Primer on Measurement Invariance Testing, *Chemistry Education Research and Practice,* **21,** 969-988.
36. Rovers, S. F. E., Clarebout, G., Savelberg, H. H. C. M., Van Merriënboer, J. J. G., (2018), Improving student expectations of learning in a problem-based environment, *Computers in Human Behavior,* **87,** 416-423.
37. Satorra, A., Bentler, P. M., (1994), Corrections to test statistics and standard errors in covariance structure analysis, In A.von Eye & C. C. Clogg (Eds.), Latent variables analysis: Applications for developmental research (pp. 399–419). Sage Publications, Inc.
38. Schmidt-McCormack, J. A., Muniz, M. N., Keuter, E. C., Shaw, S. K., Cole, R. S., (2017), Design and implementation of instructional videos for upper-division undergraduate laboratory courses, *Chemistry Education Research and Practice,* **18,** 749-762.
39. Schweizer, K., (2010), Some guidelines concerning the modeling of traits and abilities in test construction, *European Journal of Psychological Assessment,* **26,** 1, 1-2.
40. Stains, M., (2022), Keeping Up-to-Date with Chemical Education Research Standards, *Journal of Chemical Education,* **99,** 6, 2213–2216.
41. Taber, K. S., (2018), The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education, *Research in Science Education,* **48**, 1273–1296.
42. Wang, C., Cavanagh, A. J., Bauer, M., Reeves, P. M., Gill, J. C., Chen, X., Hanauer, D. I., Graham, M. J., (2021), A Framework of College Student Buy-in to Evidence-Based Teaching Practices in STEM: The Roles of Trust and Growth Mindset, *CBE Life Science Education,* **20,** 4, 1.
43. Weaver, G. C., Russell, C. B., Wink, D. J., (2008), Inquiry-based and research-based laboratory pedagogies in undergraduate science, *Nature Chemical Biology*, **4**, 10, 577-580.
44. Worthington, R. L., Whittaker, (2006), Scale Development Research: A Content Analysis and Recommendations for Best Practices, *The Counseling Psychologist,* **34,** 6.
45. Zeng, B., Wen, H., Zhang, J., (2020), How Does the Valence of Wording Affect Features of a Scale? The Method Effects in the Undergraduate Learning Burnout Scale, *Frontiers in Psychology,* **11.**