Validation of an Assessment for Introductory Linear Algebra Courses

Muhammad Qadeer Haider
Florida State University

Research-based and validated open-ended assessments are useful tools to explore students' reasoning and understanding of a subject. The primary goal of this study is to validate an assessment which can accurately measure students' conceptual understanding of four focal topics, typically covered in an introductory linear algebra course; span and linear independence, systems of linear equations, linear transformations, and eigenvalues and eigenvectors. I used the assessment data of 255 students, from nine linear algebra classes at eight different institutes across the country to validate the assessment. By administering the assessment in their classes, linear algebra instructors can gauge their students' conceptual understanding of linear algebra concepts and can identify the concepts which are generally vexatious for students.

Key Words: Assessment Validation, Linear Algebra, Conceptual Understanding

I draw on the research that has developed and validated assessments of student understanding at the undergraduate level in the areas of physics, calculus, and abstract algebra to inform the process of validation for the linear algebra assessment. A review of the literature on assessment development and validation has revealed that there are few research-based instruments available for the assessment of students' reasoning and conceptual understanding in undergraduate mathematics courses. Additionally, there is no reliable instrument available for large-scale usage which can measure students' conceptual understanding of linear algebra.

Math education literature privileges conceptual understanding of mathematics and identifies a disconnect between students' conceptual understanding and their ability to follow a procedure to produce correct answers. Students should learn mathematics with conceptual understanding and they should actively build new knowledge from their prior experience and knowledge (NCTM Principles and Standards for School Mathematics, 2000). Understanding mathematical concepts are critical in advanced mathematics but not trivial (Melhuish, 2015).

> "Conceptual knowledge is rich in relationships. It can be thought of as a connected web of knowledge, a network in which the linking relationships are as prominent as the discrete pieces of information…Procedural knowledge consists of rules or procedures for solving mathematical problems. It is also a familiarity with the individual symbols system and with the syntactic conventions for acceptable configurations of symbols (Hiebert & Lefevre, 1986, pp. 3,7)."

However, attending to procedures is not only about memorizing the list of steps to solve problems (Star, 2005; Hassenbrank & Hodgson, 2007). Researchers have also provided evidence that both methodologies of teaching concepts and teaching procedures can be integrated. Keene & Fortune (2016) have advocated the importance of the connection between teaching concepts and teaching procedures and proposed a framework for Relational Understanding of Procedures, the framework can help instructors to merge both types of teaching to help students better learn the subject matter.

The goal here is not to discuss the artificial dichotomy between procedural and conceptual knowledge but to support the argument that the connection between both is important for better learning. However, sometimes instructors can neglect the testing of conceptual understanding in regular class assessments and focus only on procedural questions (Tallman & Carlson, 2012). Therefore, assessments must be designed carefully to ensure that students have attended to the concepts in the course along with procedural efficiency.

Validated and reliable assessments are not only to assign grades, but many other important goals can be achieved through the assessment results. Teachers, principals, researchers, and organizations can use validated assessments for different purposes; for example, to provide feedback to students on their learning, to use as a diagnostic tool, to inform a placement criteria, to motivate students, to design and adjust student instructional activities, to distinguish between high and low performing students, to evaluate instructional innovations, and to sense overall performance of students, teachers, and organizations (e.g., Brown & Knight, 1994; Gibbs, 2003; Hestenes, Wells, & Swackhamer, 1992). Sainsbury and Walker (2007) argued that tests can also help students in focusing their attention and drive their learning process in the right direction. Additionally, a validated instrument can also evaluate efforts to improve learning and can help researchers to measure the quality and achievements of instructional innovations (Melhuish, 2015).

Research-based validated instruments are required to measure students' learning accurately. The central goal of this work is to validate an assessment which is sensitive to students’ ways of reasoning and understanding of linear algebra concepts.

Literature Review

In this section, I organize my summary of the literature into two main categories: studies that focus on describing different phases of assessment development and studies that focus on processes for assessment validation. Most assessment development studies also discuss validation, but the focus of the studies remain to elaborate different phases of assessment development (e.g., Carlson, Oehrtman, & Engelke, 2010; Melhuish, 2015; Sadaghiani, Miller, Pollock, & Rehn, 2013). Similarly, the validation studies also briefly describe the process of their assessment development (Barniol & Zavala, 2014; Wilcox & Pollock, 2014). Since the goal of this study is to validate the linear algebra assessment, consulting the literature on assessment validation is critically important to driving my work.

Concept inventory is another assessment development and validation approach which is gaining popularity in undergraduate STEM areas. While there is no universally accepted definition of a concept inventory, Epstein (2013) defined concept inventory as a test of a student’s most basic conceptual comprehension of a subject’s foundations, not the computational skills involved. Concept inventories measure only conceptual understanding and usually concentrate on specific topics within the course curriculum (e.g., Halloun & Hestenes, 1985a, 1985b; Hestenes & Wells, 1992; Hestenes, Wells, & Swackhammer, 1992). Although all initial work on concept inventory was in the field of physics, recently researchers have developed some concept inventories for Pre-Calculus, Calculus, and Abstract Algebra (Carlson, Oehrtman, & Engelke, 2010; Epstein, 2013; Melhuish, 2015).

It is worthwhile to do the challenging work of assessment development and validation because validated assessments produce more accurate results of students' learning than usual classroom assessments. Typically, classroom assessments are loosely structured and have several limitations including a) the instructor's expertise in the subject, b) amount of time the teacher can invest to administer, grade, and provide feedback to students, and c) performance of students in one section of a course cannot be compared with the performance of students in another section of the same course (Thissen-Roe, Hunt, and Minstrell, 2004).

Assessment validation studies usually focus on validity, reliability, and discriminatory power of the entire test and individual items on the test (Barniol & Zavala, 2014). Validity is the extent to which an instrument can measure what it is supposed to measure. To establish the validity of a test, researchers use a variety of validation techniques. Content validity is a measure of how accurately test items covered the content domain the test planned to cover, and reliability of a test is the likelihood that the test will produce consistent results repeatedly (Crocker & Algina, 2008). Cronbach's alpha is a well-known statistical method to determine how closely related a set of items are as a group. A reliability index of 0.7 and higher indicates that the test is reliable for group measures. Discriminatory power is the characteristic of a test to differentiate among high and low achievers. Some statistical analysis can also determine the quality of individual items on the test. Researchers typically determine item difficulty index and item discrimination power of individual items of assessment to validate the assessment (e.g., Barniol & Zavala, 2014; Gleason, White, Thomas, Bagley, & Rice, 2015; Wilcox & Pollock, 2014). Item discrimination is the ability of an item to differentiate between high achieving and low achieving students by establishing a relationship between how well students performed on the item and their total score on the exam (Crocker & Algina, 2008).

My review of the literature on assessment development and validation revealed that there is no valid instrument available to assess conceptual understanding of undergraduate linear algebra topics. Therefore, the purpose of this work is to validate a linear algebra assessment tool which will focus on specific concepts of linear algebra and instructors can use it in their classes.

Data Sources

Our research group has collected the assessment data as part of a broader NSF project which aimed to develop and assess a system of support for undergraduate mathematics instructors interested in teaching in inquiry-oriented ways. The project supported three subject areas: abstract algebra, differential equations, and linear algebra. The project provided participating instructors three types of support: a summer workshop, inquiry-oriented linear algebra (Inquiry-Oriented Linear Algebra IOLA; http://iola.math.vt.edu) teaching material, and a weekly online instructors' work groups (Bouhjar, Andrews-Larson, Haider, & Zandieh 2017). Previously, a team of mathematician and math educators has developed the IOLA instructional materials. The IOLA covers four major topics span, linear dependence and independence, transformations, and eigenvalues eigenvectors (Wawro, Rasmussen, Zandieh, & Larson, 2013). Linear algebra instructors usually cover these topics in introductory linear algebra classes, and the linear algebra assessment was developed to cover these four topics.

The goal of the NSF project is to improve students' learning experience in undergraduate mathematics courses; this creates a need to develop a validated assessment to measure the difference of understanding of students who attended inquiry-oriented classes. In a previous work, members of our linear algebra research group and I developed the linear algebra assessment and collected data (Haider et al., 2015). In this study, I have used the assessment data of 255 students, which were collected from linear algebra classes of nine different instructors at eight different institutions across the country to validate the assessment.

Methods of Analysis

For this study, we first need to score the assessment data. Therefore, our research group worked together to develop a reliable scoring rubric to score the assessment copies. Statistical analysis was branched into the analysis of individual items on the test and the analysis of the entire test. More details on the development of scoring rubric, scoring process, and the analysis of data are provided in the section below.

Naturally, scoring the assessment data is the first step towards the analysis of the assessment data. To maintain the reliability of the assessment, it needs a well-defined scoring rubric so different iterations of the assessment produce comparable results. I worked with three other members of our research group to develop a scoring rubric for the assessment. Initially, a senior math education researcher developed a solution key by using a variety of student approaches from pilot data. The solution key was discussed, adjusted, and explained to other members of the group to make sure that every team member completely understood every question and a potential solution of each item. I randomly selected ten copies of the assessment from the entire data set and four researchers independently identified if the given response is correct, incorrect, or partially correct according to the solution key. If we noticed other correct approaches apart from initial solution key, I added those approaches to the potential solutions.

Later, the researchers discussed and resolved any disagreements that appeared and made four categories of student responses: fully correct (awarded 3 points), partially correct (awarded 2 points), some relevant information provided (awarded 1point), and completely incorrect and irrelevant answers (awarded 0 points). For the pilot testing of the scoring rubric, I again randomly selected six different copies of the assessment from the data, made four copies of the six assessments, and every researcher in our research group scored the first question of the assessments following the scoring scheme independently, and then we compared the scores among the team members. We repeated this processes for all the questions on the assessment, and on average, there was more than 85% agreement among the researchers. We also discussed all the disagreements and came to a consensus and fine-tuned the scoring rubric accordingly.

After finalizing the scoring rubric, the next step was to score the assessment data. I randomly selected one-third of the assessment copies (i.e., 85 copies out of 255) with the help of random number generator tool. To ensure the accuracy of my scoring, I will randomly select 20% of the scores copies, and members of my research group will double code those assessment copies. We will be looking for more than 80% of intercoder reliability. The process will be repeated for the rest of the data.

Analysis of the Assessment and Validation Results

The linear algebra assessment was designed to align with the four main topics of IOLA material, which were mentioned earlier, and the goal of this study is to validate the assessment. Initial analysis shows that all the items on the test have discriminatory power and item are reasonably correlated with each other. Overall, the assessment is reliable enough to use for large groups. For the initial findings, I have scored and used 51 assessment copies in the statistical analysis. Next analysis with larger data set will support the current findings.

During the development of the test, the content validity was established through expert validation. The content areas of the assessment and selected questions for each area were consulted with three mathematics faculty members at three different institutions. The field experts helped us to identify the items which focused on the four focal topics and had potential to measure students' conceptual understanding of those topics.

I used Cronbach's alpha to measure the overall test reliability and found $\alpha = .74$ for all questions (including multiple-choice and open-ended parts), which shows that the assessment is acceptably reliable. However, when I checked the Cronbach's alpha for multiple-choice and open-ended questions separately, the values of α were dropped to .49 and .66 respectively. Item-total statistics showed that deleting any item from the test will decrease the reliability of the assessment. These statistics show that separating MCQs and open-ended items or deleting an entire item will adversely affect the assessment reliability. Statistical analysis of the assessment also revealed none of the items on the assessment have a negative correlation with other items and the corrected item-total correlation for all items is between 34% and 68%. This shows that items are not completely disconnected, but they also do not measure the same construct redundantly.

At item level analysis, the average score of all items was between 54% and 83%, which indicates that some items on the assessment were easier than others. Overall, the average score of students on multiple-choice items was 69%, and slightly lower, 65% of the open-ended items. A separate analysis of the performance of students in four focal areas showed that students performed better on the questions related to span, linear independence, and system of linear equations where the average score was above 75%. However, students were struggling with transformation and eigenvalues & eigenvectors questions where the average score was less than 65%. These results indicate that the linear algebra assessment can help to differentiate among high and low achievers and to identify the linear algebra concepts which are typically challenging for students.

Questions for Audience

- What are the methodological issues and disadvantages for having different types of questions (variety of MCQs, true/false, fill in the blanks, and open-ended) in one assessment?
- What are other appropriate validation techniques for an assessment with mixed format items?
- How can I gradually shift this work towards concept inventories? What could be possible methodological difficulties in the shift?

References

Barniol, P., & Zavala, G. (2014). Test of understanding of vectors: A reliable multiple-choice vector concept test. *Physical Review Special Topics-Physics Education Research*, *10*(1), 010121.

Bouhjar, K., Andrews-Larson, C., & Haider, M. (2017 submitted). Examining Students' Procedural and Conceptual Understanding of Eigenvectors and Eigenvalues in the Context of Inquiry- Oriented Instruction. Manuscript submitted for publication, 18 pages.

Brown, S., & Knight, P. (1994). Assessing learners in higher education. Psychology Press.

Carlson, M., Oehrtman, M., & Engelke, N. (2010). The precalculus concept assessment: A tool for assessing students' reasoning abilities and understandings. *Cognition and Instruction*, *28*(2), 113-145.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. *Holt, Rinehart and Winston*, 6277 Sea Harbor Drive, Orlando, FL 32887.

Engelhardt, P. (2009). An Introduction to Classical Test Theory as Applied to Conceptual Multiple-Choice Tests in Getting Started in PER, Vol. 2,

Epstein, J. (2013). The calculus concept inventory-measurement of the effect of teaching methodology in mathematics. *Notices of the American Mathematical Society*, *60*(8), 1018-1027.

Gleason, J., White, D., Thomas, M., Bagley, S., & Rice, L. (2015). The Calculus Concept Inventory: A Psychometric Analysis and Framework for a New Instrument. In *Proceedings of the 18th Annual Conference on Research in Undergraduate Mathematics Education* (pp. 135-149).

Haider, M., Bouhjar, K., Findley, K., Quea, R., Keegan, B., & Andrews-Larson, C. (2016). Using student reasoning to inform assessment development in linear algebra., *19th Annual Conference on Research in Undergraduate Mathematics Education* (pp. 163-177). Pittsburgh, PA. Retrieved from http://sigmaa.maa.org/rume/RUME19v3.pdf

Halloun, I. A., & Hestenes, D. (1985a). The initial knowledge state of college physics students. American journal of Physics, 53(11), 1043-1055.

Halloun, I. A., & Hestenes, D. (1985b). Common sense concepts about motion. American journal of physics, 53(11), 1056-1065.

Hestenes, D., & Wells, M. (1992). A mechanics baseline test. The Physics Teacher, 30(3), 159–166.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. The Physics Teacher, 30(3), 141-158.

Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural knowledge for teaching on student achievement. Hiebert, J. Ed. Conceptual and procedural knowledge: the case of mathematics. Hillsdale, NJ: Erlbaum, 1-27.

Keene, K., & Fortune, N. (2016). A framework for integrating conceptual and procedural understanding in the first two years of undergraduate mathematics. [Web log post.]. American Mathematical Society, Blog On Teaching and Learning Mathematics.

Melhuish, K. (2015). Determining what to assess: a methodology for concept domain analysis as applied to group theory. In Proceedings of the 15th Annual Conference on Research in Undergraduate Mathematics Education (pp. 736-744).

National Council of Teachers of Mathematics (Ed.). (2000). Principles and standards for school mathematics (Vol. 1). National Council of Teachers of Mathematics.

Thissen-Roe, A., Hunt, E., & Minstrell, J. (2004). The DIAGNOSER project: Combining assessment and learning. Behavior Research Methods, 36(2), 234-240.

Sadaghiani, H., Miller, J., Pollock, S., & Rehn, D. (2013). Constructing a multiple-choice assessment for upper-division quantum physics from an open-ended tool. arXiv preprint arXiv:1308.4226.

Star, J. R. (2005). Reconceptualizing procedural knowledge. Journal for research in mathematics education, 404-411.

Tallman, M. A., & Carlson, M. P. (2012). A characterization of Calculus I final exams in US colleges and universities. In The 15th annual conference on Research in Collegiate Mathematics Education. Portland, OR.

Wawro, M., Rasmussen, C., Zandieh, M., & Larson, C. (2013). Design research within undergraduate mathematics education: An example from introductory linear algebra. Educational Design Research—Part B: Illustrative Cases, 905-925.

Wawro, M., Rasmussen, C., Zandieh, M., Sweeney, G. F., & Larson, C. (2012). An inquiry-oriented approach to span and linear independence: The case of the magic carpet ride sequence. *PRIMUS*, *22*(8), 577-599.

Wilcox, B. R., & Pollock, S. J. (2013). Multiple-choice assessment for upper-division electricity and magnetism. arXiv preprint arXiv:1308.4619.