Formative Evaluation of Assessment Instruments for Statics

Sean St.Clair and Nelson Baker
Georgia Institute of Technology

Abstract

This paper describes a formative study that took place within the context of a larger project investigating the effects of technology on knowledge retention. In the larger project, students were evaluated at various points in time to assess their levels of learning and retention. The purpose of the formative study was not to assess students, but to evaluate the pretests, posttests, and examination questions that were later used to assess students. These instruments were tested for usability, reliability, and validity. In the formative study, the instruments were completed by students in two sections of a sophomore level mechanics course. The resulting data were compared via standard statistical techniques and the instruments were found to be reliable. The data were also analyzed for evidence of criterion-related validity and the instruments were found to be highly valid. After some changes were made based on student responses, the instruments were also found to be usable. This paper describes the formative study and the findings on the usability, reliability, and validity of the assessment instruments.

Introduction

A longitudinal study has been conducted at the School of Civil and Environmental Engineering at Georgia Tech to determine the effects of technology use in the classroom on long-term retention. Specifically, students in three sections of a statics course used two different software titles during the truss analysis portion of the course to reinforce classroom instruction[1]. The students were assessed prior to using the software and at various points in time after using the software to determine the effects of software use on learning and retention. Three different assessment instruments were used in these evaluations: a pretest, a posttest, and an examination question.

One component of the longitudinal study was to evaluate these assessment instruments to ensure that they were usable, reliable, and valid. This evaluation took place during a formative phase of the study and was conducted prior to gathering the student data on learning and retention. The formative study is the focus of this paper. The summative results on learning and retention will be presented elsewhere.

Objectives

Walker explained the distinction between formative and summative assessments in his popular Evaluation and Assessment Primer[2]. Summative assessments are performed at the conclusion of an intervention to determine the ultimate results of that intervention. Formative assessments are conducted during the course of the intervention and are intended to improve some aspect of the study. The purpose of the formative study described herein was to improve the assessment instruments that were later used to gather summative data.

In addition to improving the instruments, the formative study was also intended to provide evidence that the tests were usable, reliable, and valid. It is recommended that all research measures should be evaluated for reliability and validity[3]. The reliability of an instrument is its degree of consistency. If a student completes a test on multiple occasions and gets identical results, the test is perfectly reliable. Validity, on the other hand, is a measure of accuracy. A valid test accurately measures what it was designed to measure. Reliability and validity cannot be assumed, however. Evidence must be provided to ensure that instruments are reliable and valid. This evidence assures that the instruments accurately and consistently measure what they have been designed to measure. Additionally, when assessments are completed by students there must be evidence to show that the instruments are usable. An instrument is usable when there is no confusion or misunderstanding about how to complete it.

The objectives of the formative study stemmed from these two purposes. The two objectives that the formative study was designed to fulfill are listed below. The assessment instruments, the study method, the results, and the conclusions are described in the following sections.

- To ensure that the pretest, posttest, and examination question were usable, reliable, and valid assessment instruments.
- To improve aspects of the assessment instruments that negatively affected their usability, reliability, and validity.

Instruments

The assessment instruments were developed with input from seven different statics instructors. The tests were designed to assess various cognitive levels, from the rote memorization of truss assumptions to the qualitative evaluation of truss behavior. While it is outside the scope of this paper to describe the entire development process, summaries of the instruments are presented here. Three different tests or problem sets were evaluated in the formative study: a pretest, a posttest, and an examination question. The complete instruments are not included in this paper but they can be found in their entirety at http://epitome.ce.gatech.edu/asee.

The pretest was designed to test students' knowledge of information they needed to analyze trusses. The pretest consisted of questions on trigonometry, vector resolution, and equilibrium of forces and moments. These topics were identified by instructors of the course as prerequisite knowledge for analyzing and understanding trusses. In the summative phase of the assessment, the pretest was administered just prior to the intervention. Furthermore, the pretest was designed to be a tool to compare the different sections of statics that were involved in the summative study. The results of the pretest were compared in summative assessments to assure that all sections entered the truss analysis portion of the course with the same prerequisite knowledge.

The posttest was designed to assess students' knowledge of trusses and truss analysis. The posttest consisted of questions on truss assumptions, truss analysis techniques, quantitative truss analysis, and qualitative truss behavior. In the summative phase of the assessment, the posttest was administered just after the intervention. The posttest was designed to be a measure of learning. Additionally, the posttest was administered ten and twenty-five weeks after the intervention as a measure of long-term knowledge retention. The results of the posttests were compared in summative assessments to determine if software use had an effect on learning and retention.

The examination question, shown in Figure 1, was a quantitative truss analysis question that was part of a midterm test. The formative study was conduced in two sections of statics. The instructor of these two sections allowed the examination question to be included as part of the formative assessment. The scores on the examination question were used as a benchmark for criterion-related comparisons of validity. The role of the examination question in the validation process is discussed in the results section of this paper.

The truss below is supported by a pin at H and a roller at R (simply supported). Determine the forces in the highlighted members *GH*, *GR* & *QR*. Indicate whether the members are in tension or compression. Show all work.

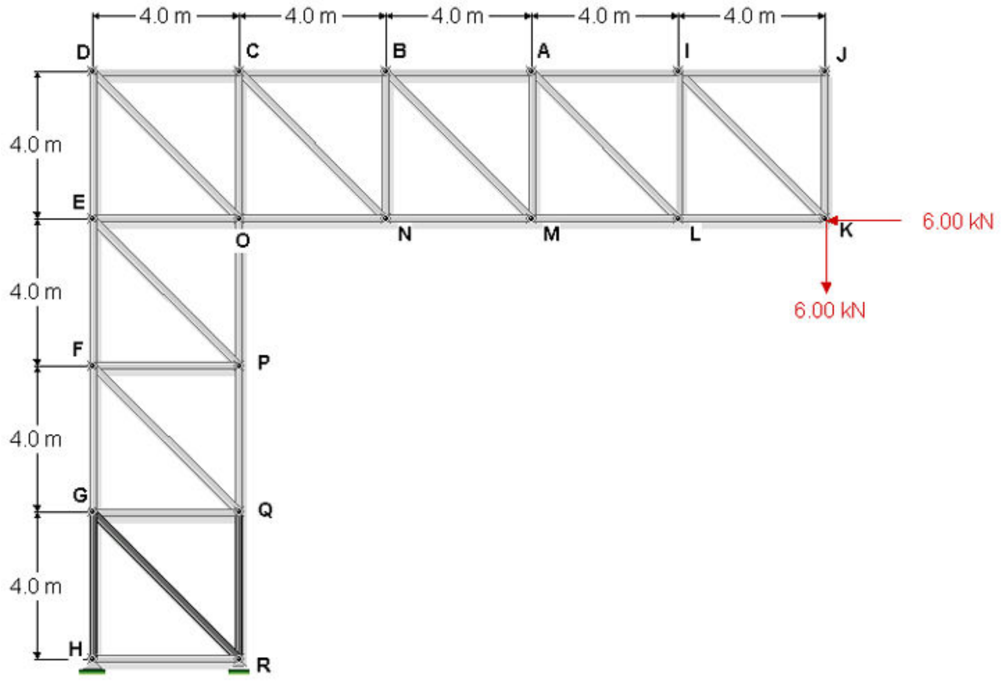


Figure 1 Examination Question

Method

The purpose of the formative study was not to assess the students, but to assess the instruments that would later be used in the summative phase of the research. As such, no educational intervention took place during this initial study. Two sections of statics participated, but neither of them used any type of software in their study of trusses. Furthermore, since the same instructor taught both sections, it was assumed that both classes received the same lectures and instructional materials.

The pretest was given to the students as a take-home assignment before the truss portion of the course. The students were informed that the test would not be graded but that they would receive bonus points for completing it. The take home test was given on a Friday and students were asked to return it the following Monday in order to receive a 1% bonus added to their final grades. Students were allowed to take as much time as they needed on the test but were asked to record how much time they spent on the assignment. Despite the fact that they were not being graded on the pretest, students were asked to try their best on each question. To avoid having the results skewed by students merely guessing the correct answers, they were encouraged to answer *I don't know* to any problem that they were unable to answer or complete. Furthermore, students were instructed to not use books, notes, or any other resource besides a calculator to answer the questions.

After the instructor completed the lectures on truss analysis, students were given the posttest. The intended methodology was to have the entire posttest administered and completed in-class. However, time constraints caused by a short summer semester only allowed for 20 minutes of class time to gather posttest data. Unfortunately, this was not enough time to administer the full instrument. Thus, one section was given half of the posttest and the other section was given the other half.

Finally, the students' truss knowledge was formally assessed via one question that appeared on their midterm examinations. The midterm test was a timed examination that was completed in class. One question on the examination was a quantitative truss analysis problem. This question was graded by one of the project researchers, as were all of the assessment instruments. Because partial credit was given on the quantitative analysis problem, a strict predefined grading rubric was designed and implemented to ensure that each student was evaluated in the same manner.

Results

The results of the various tests are presented here. The results are divided into three sections, one for each of the assessment instruments. Usability, reliability, and validity analyses of the data are also presented in this section. The results of the examination question are presented first because they were a benchmark against which the pretest and posttest data were compared.

Examination Results

All of the students, 14 from one section and 29 from the other, completed the midterm examination. The distribution of students' scores on the truss analysis question is depicted in Figure 2. The distribution is clearly bimodal, which is interesting considering that scores on the pretest and posttest were somewhat normal. A group of statics instructors, however, did not find this unusual. Truss analysis is a fairly straightforward process that students either do or do not know how to do. One of the professors referred to truss analysis questions as "light switch problems", noting that students are either on or off. As can be seen in the figure, many of the students did not even earn half of the credit available for this problem. Clearly there was room for improvement, and it was agreed among the statics instructors that an intervention involving educational technology could help the students who were otherwise confused by truss analysis.

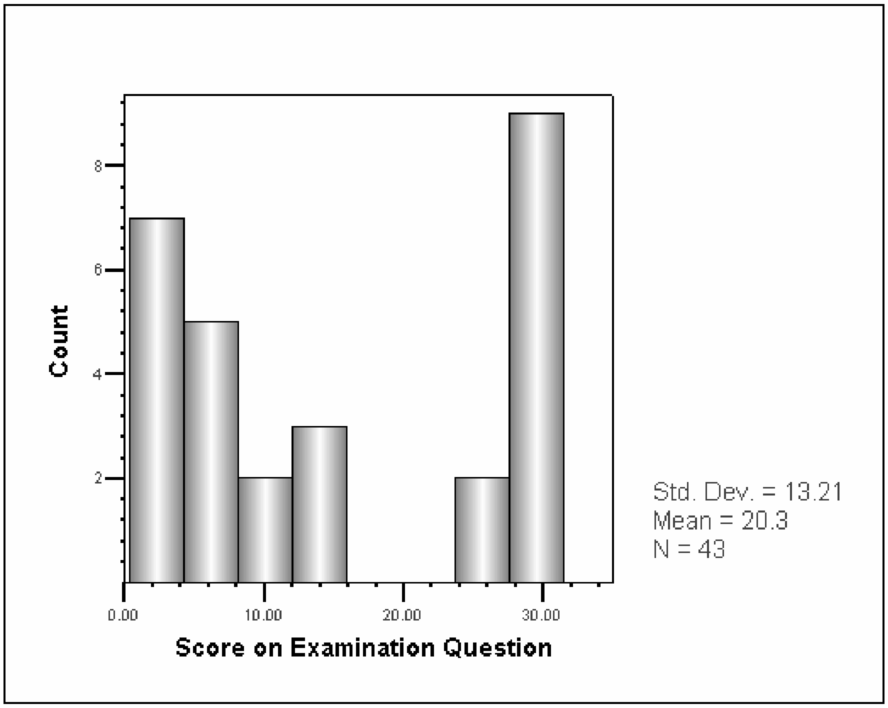


Figure 2 Histogram of Examination Question Scores

Usability of Examination Question

There were no usability issues with the examination question. The students were able to complete the problem without any concerns or confusion about the question wording or the problem diagram. As such, the problem was considered usable.

Reliability of Examination Question

The reliability of the truss analysis question shown in Figure 1 was not officially analyzed. A formal analysis of reliability requires that either a measure is given to subjects on multiple occasions (reliability across time) or that a measure consists of multiple items (reliability among items). Unfortunately, the examination question was not administered to students on multiple occasions nor did it consist of multiple items—it was just one question. As such, no formal assessment was conducted. This was acceptable because the examination question was not designed to be an assessment instrument for use in the summative phase of the research; it was simply a tool that was used to validate the pretest and posttest.

Validity of Examination Question

The examination question was validated through the use of expert validation. Expert validation occurs when a group of experts agrees that a measure accurately and completely measures a trait or ability[3]. In this case, statics instructors were recruited as experts and provided input into the design of the examination question. When the question was completed, it was presented to this group of instructors who all agreed that the instrument accurately tested students' truss analysis abilities. Thus, it was concluded that the question was valid.

Pretest Results

Six out of 14 students from one section and 15 out of 29 students from the other chose to complete the take home assignment. The pretest scores, shown in Figure 3, were out of a possible 34 points. The distribution was somewhat normal with a mean of 19 points (55.8%) and a standard deviation of 7 points (20.8%). A broad range of student abilities was clearly present, revealing that different students came into the truss portion of Statics with different prior skills and knowledge bases.

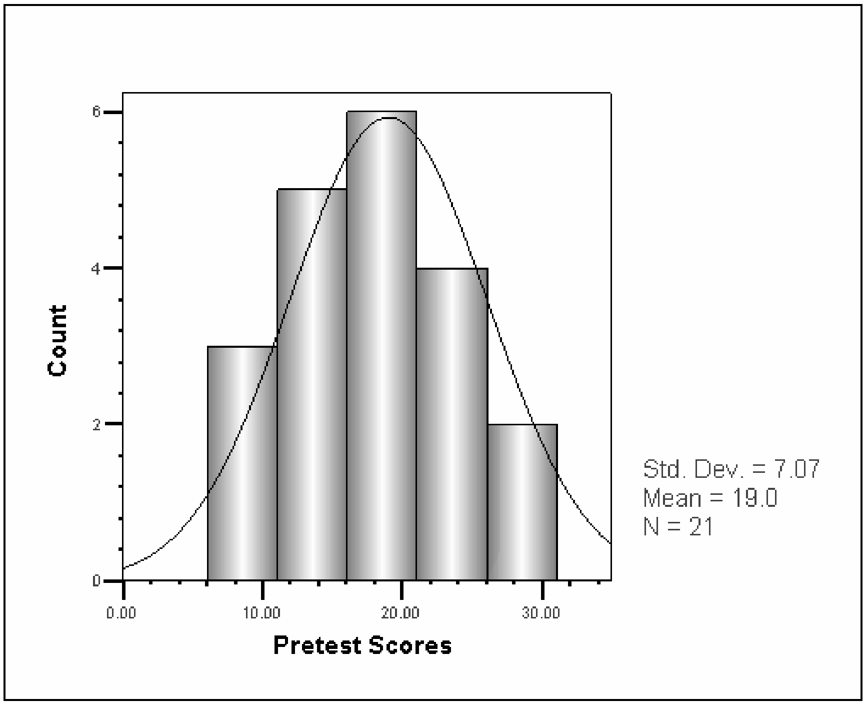


Figure 3 Histogram of Scores on Pretest Questions

Usability of Pretest

Usability was not quantitatively evaluated; it was evaluated through observations of completed instruments as well as observations of and comments from students while completing some of the instruments. A few usability issues were brought up and corrected during the formative study. Most of the corrections involved revising question wording, adding questions, or changing the format of the instruments.

One general revision was that on future assessments, students would no longer be encouraged to answer *I don't know* to problems they were not able to complete. It was observed from these formative results that, despite asking students to try to complete the problem before answering *I* *don't know*, many of them used this option as an excuse to avoid some of the more difficult questions. It was assumed that students would put more effort into each of the problems if this option were removed. Observations from summative assessments supported this assumption.

Specific concerns with some of the pretest problems were revealed through observations of the completed pretests. A number of problems were consistently missed, which is acceptable if the problem had been designed to be challenging. One such problem asked the students to solve for the reactions of a Howe truss placed and loaded on an incline. Students had studied equilibrium and should have been able to solve for reactions, but the problem appeared intimidating and was challenging. For these reasons, it was expected that most students would not attempt or successfully complete the problem. Many students, however, missed other problems that they should have been able to answer with ease. This poor performance led to revisions of two questions.

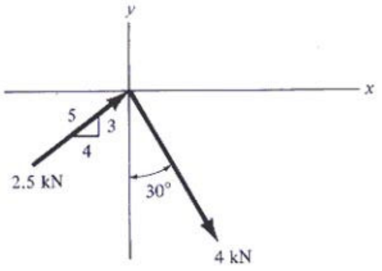
The first of these two questions initially read as follows:

> *How many reaction forces do the following types of supports provide?*
>
> *Roller* _____
>
> *Pin* _____
>
> *Rocker* _____
>
> *Fixed* _____

In a meeting with faculty members, it was agreed that the reason that many students missed this question might have been the use of the phrase *reaction forces* because any force can be broken down into any number of component forces. A rocker, for example, provides only one reaction. When placed at an angle, however, this reaction is often broken down into Cartesian components, which may be confused as being two independent reaction forces. To alleviate this concern, the wording of this question was revised for future use to read: *How many* ***unknowns*** *are associated with the following types of supports?*

The second problem to be revised initially read as follows:

> *Resolve the following force vectors into their x and y components. Add the vectors and determine the magnitude and direction of the resultant force.*



(From Hibbeler[4])

Many students missed points on this problem because they did not complete each of the steps required in the problem statement. To correct this concern, the problem was revised by breaking it down into a four separate problems, one problem for each step in the original question. In addition to these revisions, faculty members requested that a question on cross products be added to the pretest. As this was another skill students may use in truss analysis, a simple question on the topic was added for future use.

Reliability of Pretest

The pretest consisted of a number of different questions on math and equilibrium topics. Because reliability of a single instrument containing multiple items was desired (as opposed to reliability across instruments or across time), the split-half measure of internal consistency was chosen[3] to analyze the reliability of the pretest. To complete this statistical test, the pretest questions were divided into two parts. The responses to these split parts were then compared, via the Guttman Split-half method, as if they were two different sets of questions[5]. The method returned a reliability coefficient of 0.772, which exceeds the recommended minimum coefficient of 0.7[3]. It was concluded that the pretest was internally consistent and was thus reliable across time as well[3].

Validity of Pretest

Evidence of both content-related and criterion-related validity was provided for the pretest. The content-related evidence was provided via expert validation. Again, a group of statics instructors agreed that the pretest accurately and completely measured the information that students must know in order to analyze trusses. Criterion-related evidence was the degree to which a measure was related to some other measure or criterion[3]. In this case, the pretest was validated by comparing the results of pretest to the students final grades and to their scores on the examination question. Because final grades are ordinal data, a nonparametric statistical method was used. A Spearman rank order comparison was made between the three measures with a Bonferroni adjustment for two comparisons. Both correlations were significant at the 0.05 level (0.025 after the Bonferroni adjustment). The correlation between the pretest scores and the final grades had a coefficient of $\rho=0.538$ ($p<0.025$). The correlation between the pretest scores and the examination scores had a coefficient of $\rho=0.548$ ($p<0.025$). Based upon these results combined with the evidence of expert validation, it was concluded that the pretest was valid.

Posttest Results

As the method section of this paper explains, the posttest was not administered in its intended form. As one section of class only took half the test, and the other section took the other half, total scores on the posttest were not obtained. Reliability and validity analyses were conducted on individual portions of the posttest, however, and the results are presented here.

Usability of Posttest

Though the students had no concerns or questions related to the completion of the pretest, observations of their results led to a few revisions of specific posttest questions. One set of questions that were revised related to a truss with unknown dimensions. Load directions were given but the magnitudes of the loads were not. Based on this minimal, yet sufficient, amount of information, students were asked to determine whether specific members were in tension, in compression, or zero force members. This set of questions was considered important by the group of faculty members who teach the course because it asks students to think qualitatively about the truss as a whole rather than crunching numbers around a joint or section. Student responses, however, revealed that many did not take these problems seriously. The mean score on this set of problems was 1.14 out of 3 (38%). Additionally, few of the students' tests showed any scratch work, sketches, or notes on these problems. Based on these observations, it was assumed that the students probably did not spend much effort on this set of questions and may have simply guessed on them. To alleviate this problem, the revised questions asked students to not only identify the member type (i.e. compression, tension, or zero-force member) but to explain their choice in short answer form. Other than these concerns raised by the statics instructors, the posttest was completed without confusion and was thus considered usable.

Reliability of Posttest

A reliability analysis was performed on the portions of the posttest that were completed by the students in separate sections. As with the pretest data, a Guttman Split-half analysis was conducted on the posttest scores. The analysis yielded a reliability coefficient of 0.835, which was in excess of the recommended value of 0.7[3]. Based on this result, it was concluded that the posttest was reliable.

Validity of Posttest

The same techniques used to validate the pretest were used to validate the posttest as well. The group of statics instructors again agreed that the posttest was designed to accurately and completely assess a students ability to analyze trusses. Evidence of criterion-related validity was also found by comparing portions of the posttest to the scores on the examination question. This comparison yielded a highly significant correlation between the posttest and examination scores (ρ=0.891, p=0.00002). It was thus concluded that the posttest was a valid assessment instrument.

Summary and Conclusions

The formative study was successful at accomplishing the objectives of evaluating and refining the instruments that were later be used in the context of a larger, summative study. It was concluded from this study that the examination question was usable and valid. It was further concluded that the pretest and posttest were usable and reliable. When the pretest and posttest scores were compared to the examination scores, they were also found to be highly valid. Some usability issues discovered in the formative study led to minor revisions of pretest and posttest questions.

As the pretest results were not only fairly well distributed but also significantly correlated to both final grades and examination scores, it was concluded that the pretest would act as an appropriate tool for comparing the abilities of students in future sections. More specifically, the pretest was used in the summative phase of the study to ensure that various student groups possessed similar amounts of prior knowledge. Similarly, the posttest was significantly correlated to the students' examination scores and was thus appropriately measuring what it was designed to measure. Furthermore, the pretest and the posttest proved to be reliable measures, suggesting that they could be used effectively in future assessments with different study populations.

It is recommended that whenever assessment instruments are to be used in a research project, the instruments should be evaluated to ensure that they are usable, reliable, and valid. If the instruments are not properly designed and evaluated, the results that stem from them may not be accurate or applicable. In order to ensure that a research project is valid and that the results of the project can be applied to new and unique situations, the instruments must be properly validated. An example of such validation has been provided in this paper, which documents the formative evaluation of assessment instruments to ensure that they were reliable, valid, and usable.

References

[1] St.Clair, S.W. & Baker, N.C. (2003). Pedagogy and technology in statics. *Proceedings of the 2003 American Society for Engineering Education Annual Conference and Exposition*, Session 2793.

[2] Walker, N. (1995). Evaluation and Assessment Primer, available WWW: http://www.succeed.ufl.edu/content/Evaluation/primer.html, accessed Feb, 2004.

[3] Whitley, G.E. (1996). Principles of Research in Behavioral Science, Mayfield Publishing Company, Mountain View, CA.

[4] Hibbeler, R.C. (2001). Companion Website, WWW: http://cwx.prenhall.com/hibbeler, accessed June, 2002.

[5] SPSS Inc, (1999). SPSS Base 9.0: Applications Guide.

Biographies

SEAN W. ST.CLAIR
Sean St.Clair is a Ph.D. candidate in the School of Civil and Environmental Engineering at Georgia Tech. He received a B.S. in civil engineering from Utah State University and a M.S. in civil engineering from Georgia Tech. Under the advisement of Dr. Nelson Baker, he has conducted numerous projects in the areas of engineering education and educational technology and assessment. Sean can be reached at sean.stclair@ce.gatech.edu.

NELSON C. BAKER
Dr. Nelson Baker is the Associate Vice Provost for Distance Education at Georgia Tech and an Associate Professor in the School of Civil and Environmental Engineering. He is a nationally recognized leader in educational technologies designed for and applied to engineering having won numerous awards in this area. Dr. Baker's research has led to the development and assessment of unique technologies such as virtual reality interfaces for education.