Formative Assessment: An Illustrative Example Using "Alice"

Ashlyn Hutchinson[a], Barbara Moskal[a], Wanda Dann[b], Stephen Cooper[c]
Colorado School of Mines[a]/ Ithaca College[b]/ Saint Joseph's University[c]

Abstract

There are two primary phases of assessment: formative and summative. The formative phase of assessment focuses upon improving the study's design, methodologies and evaluations as the study is being implemented. Summative assessment, on the other hand, is used to evaluate the overall effectiveness of the research intervention. The appropriateness of the conclusions that are drawn during the summative phase is partially dependent on the formative phase. Difficulties that arise with respect to data collection and faulty instruments damage the validity of a study's final results. These problems can be rectified during the formative phase by carefully selecting and/or creating assessment instruments and conducting a pilot investigation before summative data is collected. This paper illustrates the formative phase of assessment and how the information collected during the formative phase was used to improve the design of a study that investigated an innovative approach to teaching introductory computer science. This work was supported in part by NSF-03020542.

I. Introduction

In most reported engineering education studies, the emphasis of the discussion concerning assessment is based on summative data. Summative data, after all, allows the researcher to examine the extent to which the stated goals and objectives of the investigation have been reached. The formative phase of the assessment process is often not reported, yet the validity of the conclusions drawn from an investigation is partially dependent on this phase. Formative assessment helps the researcher to improve the design and implementation of a project while the project is underway.[1] It is during the formative phases of assessment that initial validity evidence is collected and analyzed.[2] During this critical stage of the assessment process, changes can be made to the research and assessment design and these changes can improve the quality of the information that is obtained during the summative phases of assessment. This paper directly illustrates the formative phase of assessment and how information acquired through this phase was used to improve the *Java-based Animation: Building virtual Worlds for Object-oriented programming in Community colleges* (JABRWOC) project and its assessment.

The JABRWOC project is a three year project, funded by the National Science Foundation (NSF), which aims to improve the community college approach to computing education (NSF, DUE-0302542).[3] JABRWOC has the following three goals: 1) to combat high attrition levels in first year computer science and information technology courses, 2) to strengthen the appeal of computer science, thereby increasing the number of computer related majors, and 3) to introduce a much needed programming component into computer literacy classes.[3] The JABRWOC research team seeks to reach these goals by developing and testing curricular materials that introduce programming to community college students using a unique, multimedia based object oriented programming environment called Alice. Alice was developed by Dr. Randy Pausch, Carnegie Mellon Institute of Technology and is freely available on-line.[4] Drs. Cooper and Dann have developed a curriculum to accompany the Alice software, henceforth known as the Alice curriculum, which includes a textbook[5] and classroom implementation guidelines.[6] The projects' assessment efforts throughout the academic year 2003-2004 were to research, select, develop and validate reliable assessment instruments that were aligned with the project goals.

II. Research Question

As was previously discussed, the primary focus of this paper is to provide an illustrative example of how formative assessment may be used to improve the design and assessment of a research investigation in education. This has been restated in the following research question:

> How has formative assessment been used to improve the design and assessment of the JABRWOC research project?

III. Methods

Throughout the academic year 2003-2004, the JABRWOC project assessment team's efforts were focused on formative assessment. Specifically, the assessment team sought to develop or select valid and reliable assessment instruments that were aligned with the project goals, test these instruments through a pilot investigation and use the results of these pilots to improve the project's design and its assessment. This section describes the selection of instruments and the process for piloting these instruments.

A. Selection of Instruments

Based on the goals of the project, the research team determined that there were two student constructs that needed to be measured as part of the JABRWOC project. These were students' attitudes towards computer science and students' knowledge with respect to key concepts within computer science. This led to the decision that two assessment instruments would be needed: a computer science attitudes survey and a computer science content assessment.

During the fall of 2003, Ashlyn Hutchinson, a graduate student in the Mathematical and Computer Sciences Department at the Colorado School of Mines, began to investigate appropriate attitude surveys. Using the ERIC and ETS Test Link search engines, she completed a search of attitude survey abstracts.[7,8] This search resulted in the identification of 22 assessment instruments. Each of these abstracts was reviewed to determine whether it could be appropriately used with a community college population. This further limited the selection to three instruments. These instruments were then reviewed by the project investigators and the Loyd-Gressard Computer Attitude Scale was chosen.[9] A major benefit of using this instrument was that a great deal of validity evidence had already been collected by the original developers, reducing validity concerns within the current investigation. In addition, the Loyd-Gressard survey employs a Likert rating scale. Respondents are generally familiar with this type of rating scale, and it is an effective method to collect attitudes. By using a Likert rating scale, respondents have the advantage of taking a survey in a format familiar to them. The study benefits from the fact that this survey format is easy to complete; respondents are likely fill out the entire survey, giving the study complete data with which to work.[10,11]

Next, the Alice Content exam was developed. Through discussions with Drs. Pausch (developer of the Alice software) and Drs. Cooper and Dann (developers of the Alice curriculum), the evaluator, Dr. Moskal, determined that no appropriate computer science concept exams were available in the literature that could be used to measure students' knowledge with respect to the Alice curriculum. Therefore, such an exam needed to be created.

The investigators and the evaluator worked together to create a list of student outcomes that were expected to result from completing the Alice curriculum. Based on these outcomes, Cooper and Dann created a preliminary list of multiple choice questions. Hutchinson reviewed these questions using literature on how to create effective multiple choice questions as a guide.[12,13] She then discussed any identified problems that resulted from this review with Cooper and Dann. Jointly, Cooper, Dann, Moskal and Hutchinson revised the multiple choice assessment to be consistent with the literature.

B. Pilot Testing

Three community colleges participated in the pilot JABRWOC study. They were: Camden County College (CCC) in New Jersey, the Community College of Philadelphia (CCP) in Pennsylvania, and Tompkins Cortland Community College (TC3) in New York. In order to study the effects of the Alice curriculum, data was collected from two groups: course sections implementing the Alice curriculum in classrooms (treatment group) and course sections learning similar concepts but not offering instruction with Alice (control group). Students from both treatment and control groups provided signed consent to participate in this study.

Both treatment and control groups filled out a demographics survey at the start of the study. On the demographic survey, respondents reported their age, gender, ethnicity, year in school, major, and background in computing. Respondents also identified which classes they were enrolled in for the current semester.

Both the treatment and control groups completed the Attitudes Survey at the beginning and end of their computer science course. Attitude scores were then determined by calculating an overall composite score; a higher composite score indicates a more positive attitude toward computer science. The Alice Content exam was completed only by the treatment group at the beginning and end of the Alice course. The questions that comprised the Pre Alice Content Assessment (pretest) and Post Alice Content Assessment (posttest) were identical. The control group did not participate in this portion of the study, because the exam is based on Alice concepts. Therefore, any exam results from students who had not been instructed with the Alice curriculum would not be meaningful. Performance on the pre and posttest were calculated as the overall percentage of correct responses.

Although efforts were made to maintain consistency in data collection across all groups, it should be noted that in some of the treatment classrooms, the Alice curriculum was used throughout the semester.[14] This is in contrast to other course sections, where the Alice curriculum was used for only five weeks of the course. In the case of a five week section, the materials were administered during the first few days and last few days of the Alice portion of the course.

The data collected from the beginning of the semester and data from the end of the semester, henceforth known as Pre Data and Post Data, was first analyzed separately, and then comparisons were made. Using the statistical software MINITAB, descriptive statistics were run on both the Pre Data and Post Data to determine basic information about the population.

IV. Pilot Results

Based on the pilot investigation, a number of changes have been made in the study's design and assessment. The first section describes the revisions that have been made to limit the loss of data during the full investigation. This is followed by a brief description of how the pilot information was used to revise the assessment instruments. The final section discusses software limitations that were identified. In each situation, the remedy that has been implemented in this investigation is discussed.

A. Data Loss

The response rate from the Content Exam and Attitude Survey was analyzed. The Pre materials showed fairly high numbers. However, when the Post materials were collected, a dramatic decrease in responses became evident. As is suggested by Table 1, this is primarily accounted for by a dramatic decrease in the responses on from pre to post with respect to CCP. To better understand the drop in response rate, the evaluators interviewed a representative from CCP. Based on this interview, the evaluators learned that CCP experienced technical difficulties when administering the Post materials on the last day of class. The server that administers the Assessment and Attitudes Survey was down at the desired administration time. Since there was no other opportunity for students to complete the assessment, this caused an extreme loss of data in the pilot administration. In future administration, community college instructors will be asked to administer these instruments at least a week before the end of classes, ensuring that additional time is available if further server problems emerge.

Table 1. Response Rates

	CCC		CCP		TC3	
	Pre	Post	Pre	Post	Pre	Post
Attitude Survey	48	34	102	28	20	17
Content Exam	16	9	72	12	38	27

Another observation that was made during the pilot administration was the prevalence of duplicate entries. Two factors appeared to contribute to this problem. At TC3, students used identification numbers that exceeded the limit of the database. This resulted in some numbers losing the distinguishing digits, resulting in overlaps. More explicit instructions have since been developed to assist the schools in creating appropriate identification numbers. Second, since the instruments were administered on-line, the researchers realized that some duplicates were due to students hitting the "submit" button more than once. For the pilot data, these duplicate entries were identified and checked for completeness. The more complete entry was kept; in cases where entries were identical, one was deleted. For future studies, directions have been added requesting students only hit "submit" once.

Another contributing factor that resulted in lost data was that students were only asked to identify themselves as treatment or control on the demographics survey. Therefore, Content Exam and Attitude Survey data corresponding to identification numbers that did not fill out the demographics form had to be discarded. To avoid future confusion, a question regarding section numbers has been added to the Demographic Survey, Content Assessment, and Attitudes Survey.

An unexpected result in the pilot data was the large number of women that participated in the study from CCP. Table 2 illustrates these differences. Having more women than men enrolled in a course for computer science majors is not typical, and therefore, queries were made into the nature of this anomaly. After discussions with the head investigator and instructors at CCP, it was determined that the course CIS 103 had an unusually high number of women. This course primarily focuses on computer literacy rather than computer programming. Computer literacy courses are commonly taken by non-computer science majors. When the data was examined with CIS 103 removed, males outnumbered females, as prior research suggests is common in computer science programs. This observation was important because it raised the researchers' awareness that due to the different participating populations, computer literacy courses need to be examined separately from computer programming courses.

Table 2. Male vs. Female Responses

	Male	Female
CCC	45	4
TC3	15	10
CCP	37	47
CCP (CIS 103 excluded)	25	19

B. Instrument Revision

As was discussed, a great deal of effort was dedicated to ensuring that the instruments would result in valid data prior to the pilot study. To further establish the validity of the multiple choice instrument, a question-by-question analysis was completed on student responses. Out of 16 questions total (numbered 4-19), only two questions had a decrease in students obtaining the correct answer. These questions 8 and 19 were analyzed again, checking for ambiguities. After scrutiny of these questions, it was determined that the results were most likely due to chance, rather than faulty problems. No revisions were made.

A change that was made to the Content Exam was the addition of a set of questions that address general computer science related topics. After reviewing the first year's assessment report, the investigators determined that they wanted to be able to make comparisons between the treatment and control groups knowledge of basic computer science concepts that are not platform specific. To account for this need, general questions were added and these questions will be completed by both the treatment and control groups at the beginning and end of the investigation. The treatment group will also continue to complete the questions that are specific to Alice.

The Demographics Survey allows the evaluators to examine responses based on different student characteristics. One area of interest in this investigation is the impact that the Alice curriculum has on students of different genders and races. The original Demographics Survey did not include an option for students identifying themselves as "Multi-racial". Interviews with the head investigators for each community college indicated that many students were frustrated with the lack of options to properly identify themselves. This option has since been added to the survey.

In addition, the researchers wanted the option of examining student responses based on the students' selected majors. Originally, the Demographics Survey asked the students to indicate whether they were in Computer Science (CS), Computer Information Systems (CIS), or a non-computer field. The revised survey allows the students to identify themselves as being in Computer Science (CS), Computer Information Systems (CIS), Computer Graphics or Computer Systems Technology.

C. Software Limitations

The computers at TC3 were running Windows 98, with which the Alice software works poorly. Most instructors selected to stop using the Alice curriculum due to this problem. The facilities at TC3 have been upgraded to run Windows XP, which will eliminate those technical difficulties that TC3 users experienced with the Alice software.

D. Summary

Although further preliminary analysis was done regarding the overall effect of the Alice curriculum, the previous results allowed the evaluators to determine what changes should be made to the study design. In addition to the changes mentioned above, the co-project investigators decided to have all community college leaders and evaluators submit quarterly reports outlining the progress being made at each institution. By doing this, administrative problems can be dealt with as they arise and hopefully not adversely affect the data collection.

V. Conclusions

As the discussion above illustrates, a great deal was learned through the formative assessment of JABRWOC. By piloting the assessment instruments, the investigators and evaluator were able to make adjustments to the project and these adjustments are likely to result in increased validity during the summative phases. Changes have been made to reduce the loss of data and to improve the quality of the assessment instruments. Additionally, problems were identified with respect to the Alice software and the computer equipment that was available at a given institution. This was dealt with by replacing the equipment and recognizing that there are systems limitations when using this software and curriculum. Had the investigators selected not to complete the formative phase of assessment, many of the above discussed problems would have remained unrecognized and plagued the final investigation.

References
1. Frechtling, J., *The 2002 User Friendly Handbook for Project Evaluation*. Washington, DC: National Science Foundation (NSF 02-057), Division of Research, Evaluation and Communication, 2002.

2. Moskal, B., Leydens, J. & Pavelich, M. (2002). "Validity, reliability and the assessment of engineering education". *Journal of Engineering Education*, 91(3), 351-354. (Journal)

3. Cooper, S., Dann, W., & Moskal, B. *Java-Based Animation in Building viRtual Worlds for Object-oriented programming in Community colleges*. NSF-DUE-0302542.

4. Alice v2.ob Learn to Program Interactive 3D Graphics, http://www.alice.org (accessed December 2004)

5. Cooper, S., Dann, W., & Pausch, R. (2005) *Learning to Program with Alice Beta Version*. Prentice Hall.

6. Curricular Materials for Learning to Program with Alice, http://www.sju.edu/~scooper/alice/course/alice.html (accessed December 2004)

7. SearchERIC, http://searcheric.org (accessed October 2003)

8. ETS Test Link, http://www.ets.org/testcoll (accessed October 2003)

9. Loyd, B.H. and Gressard, C.P. Computer Attitude Scale. *Journal of Computing Research,* 15(3), 241-259.

10. Suskie, L.A. (1996) *Questionnaire Survey Research: What Works* (2nd Edition). Tallahassee, Florida: Association for Institutional Research, Florida State.

11. American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing.*

12. Kehoe, J. Practical Assessment, Research and Evaluation, *Writing Multiple-Choice Test Items, 1995,* Online: http://www.edresearch.org/scripts/seget2.asp?db=ericft&want=http://www.edresearch.org/ericdb/ED398236.htm (accessed December 2004)

13. Frary, R.B. Practical Assessment, Research and Evaluation, *More Multiple-Choice Item Writing Do's and Don'ts, 1995,* Online: http://www.edresearch.org/scripts/seget2.asp?db=ericft&want=http://www.edresearch.org/ericdb/ED398238.htm (accessed December 2004)

14. Gay, L.R. *Educational Research: Competencies for Analysis and Application (3rd Ed.).* New York: Macmillan Publishing Company, 1987.

Biographical Sketch

ASHLYN HUTCHINSON
Ashlyn Hutchinson (ashutchi@mines.edu) received her B.A. in Mathematics from the University of Colorado at Boulder, and will receive her M.S. in Applied Mathematics from Colorado School of Mines, expected May 2005. She is a Research Assistant for Dr. Barbara Moskal in the Mathematical and Compute Sciences Department at the Colorado School of Mines. Her research interests include engineering education and assessment.

BARBARA M. MOSKAL
Barbara M. Moskal (bmoskal@mines.edu) received her Ed.D. in Mathematics Education with a minor in Quantitative Research Methodology and her M.A. in Mathematics from the University of Pittsburgh. She is an Associate Professor in the Mathematical and Compute Sciences Department at the Colorado School of Mines. Her research interests include student assessment, k-12 outreach and equity issues.

WANDA DANN
Dr. Wanda Dann is an Associate Professor of Computer Science at Ithaca College. Her research has encompassed program visualization and object-oriented and event-driven programming. Dr. Dann has provided leadership in the international computer science education community, serving as SIGCSE 2004 Program co-Chair and SIGCSE 2005 Symposium co-Chair.

STEPHEN COOPER
Stephen Cooper is an Associate Professor of Computer Science at Saint Joseph's University. He taught previously at Rivier College, serving as Computer Science program director. He has also worked at IBM as a systems programmer. Dr. Cooper's research interests lie in the semantics of programming languages as well as in program visualization. He has been the principal investigator for several National Science Foundation and private grants.