Comparing Design Team Self-Reports with Actual Performance: Cross-Validating Assessment Instruments

Robin Adams[1], Pimpida Punnakanta[1], Cynthia J. Atman[1,2], Craig D. Lewis

[1]Center for Engineering Learning and Teaching
[2]Department of Industrial Engineering
University of Washington

Assessing student learning of the engineering design process is challenging. Students' ability to answer test questions about the design process or record their design activities may differ significantly from their actual performance in solving "messy" open-ended problems. In the Pacific Northwest, multi-university participants in a National Science Foundation supported project (Transferable Integrated Design Engineering Education, TIDEE) have implemented and disseminated a Mid-Program Assessment instrument for assessing engineering student design competency. One part of the instrument requires student teams to document (e.g., self-report) their design decisions and processes while engaged in a design task. These written self-reports are scored using a rubric that has demonstrated a high inter-rater reliability. We are interested in comparing the scores derived from these self-reports with measures of actual design performance. Our research method for analyzing design performance is verbal protocol analysis. In this study, eighteen teams of students (2-6 students per team) from four different institutions were videotaped as they completed the TIDEE Mid-Program Assessment. In this paper we provide 1) a description of the assessment instrument, 2) our research methods for assessing the validity of the instrument, 3) examples of comparing self-reports to performance, and 4) a summary of our findings. We conclude with a discussion of the strengths and weaknesses of this study, as well as implications for teaching and assessing engineering student design competency.

Introduction

To compete in an increasingly global economy, the education of tomorrow's engineers needs to emphasize competency in the solving of open-ended engineering design problems. This theme is evident in the growing level of collaboration among accrediting agencies, industry, and federal funding agencies to support research on the assessment of student learning and to encourage excellence in curriculum and pedagogy that provide an exposure to engineering practice[1-3]. Also, the implementation of the new ABET EC 2000 criteria[4] makes it necessary for engineering programs to identify, assess, and demonstrate evidence of design competency. These changes in accreditation have expanded a goal of assessing student learning outcomes to making judgments about curricula and instructional practices with an aim towards continual improvement.

Assessing student learning of the engineering design process is particularly challenging, and efforts to assess design competency are varied[5-6]. Examples of using surveys include self-assessments of abilities and knowledge[7-8] and peer-based instruments where students assess the competency of their peers[9-10]. Examples of performance-based assessments include: juries where

experts review and assess student work[11-12], portfolios where students record instances of work illustrating their competency[13-14], and tasks where students demonstrate their competency through written self-reports[15-16].

Each of these methods have advantages and disadvantages[17-18]. For example, surveys may be easier to administer and analyze, yet the design of effective surveys requires considerable knowledge. Also, most surveys do not provide direct measures of learning but rather only self-assessed perceptions of knowledge and ability. Self-report models of assessment have been shown to be efficient methods for assessing breadth and depth of engineering student knowledge[19] and are believed to be psychometrically adequate[20-21]. However, for complex and ambiguous design problems, students' abilities to answer questions about the design process may differ significantly from their actual performance in solving these open-ended problems in a collaborative manner.

One way to address such limitations is to use triangulation methods to cross-validate results[17,22]. In this paper we present the results of a study to cross-validate derived scores from engineering students' written self-reports of their design activity with measures of observed design performance. This includes: 1) a description of the assessment instrument, 2) our research design for assessing the validity of the instrument, 3) examples of comparing self-reports to performance, and 4) a summary of our findings. We conclude with a discussion of the strengths and weaknesses of this study, as well as implications for teaching and assessing engineering student design competency.

Validating an Assessment Instrument

In the Pacific Northwest, multi-university participants in a National Science Foundation supported project entitled Transferable Integrated Design Engineering Education* (TIDEE) have implemented and disseminated a tool for assessing mid-program engineering student design competency[15,23]. TIDEE and industry participants collaborated in a series of workshops to identify targeted engineering design outcomes. These outcomes were incorporated into an assessment instrument to assess knowledge of team-based engineering design at the mid-program level. The written assessment evolved through several iterations to become the three-part TIDEE Mid-Program Assessment instrument[23]. This instrument is available online at the TIDEE website[24].

In Part I of the Mid-Program Assessment instrument, students are given 15 minutes to respond in short essays to three written prompts. The focus of Part I is to assess individual design knowledge in three areas: a) engineering design process, b) effective teamwork, and c) effective communication. Design performance is assessed in Part II by having students respond as a team to a 45 minute structured design task and document their efforts (a self-report). Effective teamwork and communication is assessed in Part III by having individual students write a reflective essay based on their team performance in the Part II design task.

For the design activity (Part II), students are self-organized into teams and are requested to design a testing procedure for an assigned hand tool (pet nail trimmers or tree pruning shears).

* URL: http://www.cea.wsu.edu/TIDEE

Teams are provided with a hand tool specimen along with any associated product information such as material specifications, instructions for use, and any supplementary devices (e.g., extra razor blades). In addition, teams are provided with worksheets to document their design decisions and activities. Specific prompts include: defining each team member's role and responsibilities, recording the steps the teams used to complete the acitivity (a design log), listing and explaining customer expectations for the tool, identifying and justifying the most essential customer expectation, and describing a complete testing procedure to address this customer expectation. This procedure may include sketches and plans for data collection or analysis.

Our study goal is to identify the level of cross-validity between observable design behaviors and written self-reports of team design activity. Here, self-reports are defined as written documents describing team design performance and activities. These self-reports may be written prior, during, or after engaging in a performance-based design task. The process of cross-validating the design component of the TIDEE Mid-Program Assessment instrument involved two levels of analyses. The purpose of our first level of analysis was to determine 1) the existence of significant differences in our study measures across the participating institutions, and 2) the extent to which the three individual parts of the assessment instrument measured distinct competencies. If the institutions do not significantly differ we can group our samples into a single population and increase our statistical power. Otherwise, we would be need to segment our data into groups associated with each institution. In addition, if there are little or no correlations across the scores derived from each part of the instrument, we can limit our analysis to scores derived only from Part II of the instrument.

The purpose of our second level of analysis was to compare the scores derived from the self-reports with measures of observed design performance. In other words, we sought to assess the content validity of the TIDEE Mid-Program Assessment instrument by identifying the extent to which performance measures provide evidence for the derived self-report scores. Our general hypothesis is that teams that received credit for a particular element of the assessment instrument spent more time in design activities associated with that element. In addition, we expected that teams that received higher total scores progressed farther into the latter stages of the design process and transitioned more frequently across design activities. Transitioning may be described as a behavior of moving from one design activity to another. Examples include transitioning from gathering information about a design problem to evaluating the feasibility of a design solution or transitioning from modeling a design solution to revisiting the problem requirements. Measures of progression and transitioning have been associated with greater engineering experience and design success[25], and the number of transitions is believed to be a surrogate measure for design iteration[26].

Methods

Our research method for analyzing observed design performance is verbal protocol analysis (VPA). In verbal protocol analysis subjects think aloud as they perform tasks, providing the researcher with rich and detailed data that can be used to describe and empirically analyze problem solving behavior[27-28]. Verbal protocol studies have been successfully utilized to identify how designers introduce information or knowledge into the design process[30] to quantify differences in approaches[25-26,29], and to measure the effectiveness of teaching methods[31].

Examples of other methods for studying design activity, particularly for design teams, include ethnography[32] and videotape analysis[33].

Eighteen teams of students from four different institutions in the Pacific Northwest participated in this study. Institutions included both two and four year degree-granting programs and teams ranged from two to six students. Students from the two year programs were predominantly freshmen and sophomores; students from the four year programs were predominantly juniors although some were seniors. Each team completed the full Mid-Program Assessment activity. Observations of the student teams indicate few, if any, problems with completing the task on time or working in a group. Sessions from Part II (the design task) of the Mid-Program Assessment activity were videotaped. During this process students were prompted to think aloud. Videotapes were transcribed and segmented based on distinct idea units in which new segments were marked by a change in context[27,34]. Two independent coders segmented each transcript and the average reliability for this process was high (88%). All disagreements in segmenting were arbitrated to consensus.

Segments were coded by using an existing coding scheme that has been shown to be both reliable and useful in characterizing differences in design performance and design experience[25,30,35]. A description of these codes is provided in Table 1. These codes include: Problem definition (PD), Gathering Information (GATH), Generating Ideas (GEN), Modeling (MOD), Feasibility (FEAS), Evaluation (EVAL), Decision (DEC), and Communication (COM). Because teams did not implement their designs, the implementation code (IMP) was not used in this study. The average inter-rater reliability between independent coders was 85%, and all disagreements were arbitrated to consensus. Coded transcripts were imported into MacSHAPA, a software program for analyzing verbal data[36] and time stamped. Data was time stamped from the videotape to determine the amount of time associated with each coded segment. Measures of design performance generated from MacSHAPA include: 1) representations over time of design activity (e.g., timelines), 2) the amount of time spent in each design activity, and 3) the number of moves between activities (e.g., transitions). For our purposes, transition activity measured in MacShapa is used as a surrogate measure for design iteration[26].

Table 1. Design activity codes[31,35]

Abbreviation	Design Activity Code	Description
PD	*Problem Definition*	Define what the problem really is, identify constraints, identify criteria, reread problem statement or information sheets, and question the problem statement.
GATH	*Gathering Information*	Searching for and collecting needed information.
GEN	*Generating Ideas*	Develop possible ideas for a solution, brainstorm, and list different alternatives.
MOD	*Modeling*	Describing how to build an idea, how to make it, measurements, dimensions, and calculations.
FEAS	*Feasibility Analysis*	Determining workability, verification of workability, does it meet constraints, criteria, etc.
EVAL	*Evaluation*	Comparing alternatives, judgment about various options (is one better, cheaper, more accurate, etc.).
DEC	*Decision*	Select one idea or solution among alternatives.
COM	*Communication*	Define the design to others; write down a solution or instructions.
IMP	*Implementation*	Produce or construct a physical device, product, or system.

The scoring of the TIDEE Mid-Program Assessment self-reports for each team was completed by individuals trained in the scoring process[15,37]. Documents were scored "blind": scorers did not have access to the videotape sessions. Scoring was based on established decision rules[24]. Rules specific to Part II of the instrument are provided in Table 2. In this table, each TIDEE code signifies a TIDEE design element score. For example, the TIDEE code AB1 refers to the decision rule for the first question of Section AB. As shown in Table 2, teams could receive from zero to a total of 11 points. Inter-rater reliability for this scoring process was consistenly high (>80%) and differences were arbitrated to consensus.

Table 2. TIDEE scores and decision rules for Part II

TIDEE Code	TIDEE Design Element	TIDEE Decision Rule*	Credit
AB1	Section AB #1	Gathering tool information considered in design log	1
AB2	Section AB #2	Time usage considered	1
CD1	Section CD #1	> 5 ideas listed	1
CD2	Section CD #2	= 3 ideas explained	1
CD3	Section CD #3	one customer expectation selected	1
CD4	Section CD #4	selection of one customer expectation rationally justified	1
E1	Section E #1	relevant ideas for tests provided	1
E2	Section E #2	detailed procedural steps for at least one test	1
E3	Section E #3	variability/replication of test results considered	1
E4	Section E #4	quantification of test results provided	1
E5	Section E #5	criteria for passing test procedure provided	1
			11 Total

*Note: Decision rules are described in depth in reference [24].

Table 3. Mapping between TIDEE design scores and design performance measures

	Design Activity Codes[31,35]								
TIDEE Code	PD	GATH	GEN	MOD	FEAS	EVAL	DEC	COM	TRANS
AB1		**X**							**X**
AB2		**X**							**X**
CD1	**X**								**X**
CD2	**X**								**X**
CD3	**X**								**X**
CD4	**X**								**X**
E1			**X**			**X**	**X**		**X**
E2				**X**				**X**	**X**
E3					**X**				**X**
E4				**X**					**X**
E5					**X**				**X**

To compare observed design performance to students' written self-reports we generated a map relating scores in Part II of the Mid-Program Assessment instrument (TIDEE code in Table 2) to the design activity codes presented in Table 1. This map is summarized in Table 3. Because the number of transitions represent a broad measure of design performance, this measure was mapped to the cumulative total derived score. As illustrated in Table 3, the mapping was not

always exact. For some situations a single TIDEE code mapped to one design activity code (e.g., E4 and Modeling); for others, a combination of TIDEE codes mapped to a single design activity code (e.g., PD-Problem Definition). There were two instances where the derived scores for the TIDEE codes could represent multiple activities (e.g., E1 and E2). This occurred because a decision rule for this score represented multiple design activities. For example, the score for E1 involved multiple design activities associated with selecting a "relevant" test procedure. As such, teams that received this credit would be more likely to spend time generating ideas for a test procedure (GEN), evaluating these test procedures for how they addressed the primary customer expectation (EVAL), and selecting a final test procedure (DEC).

Results

As stated earlier, our efforts to cross-validate the Mid-Program Assessment instrument involved two levels of analyses. The goal of the first level of analysis was to determine the extent to which the three individual parts of the assessment instrument measured distinct competencies and the extent to which study measures significantly differed across participating institutions. The goal of the second level of analysis was to validate the scores derived from the self-reports with coded measures of observed design performance.

Comparing Across Individual Parts of the Mid-Program Assessment Instrument

Before responding to our research questions we first analyzed whether or not our data met the normality assumption. Most of our study measures were normally distributed; however, the amount of time spent in Evaluation and Decision activities were not normally distributed. One explanation may be that few teams engaged in these activities, and those who did spent less than 1% of their total design time. These were not considered to be critical violations.

Table 4. Correlations between parts of the Mid-Program Assessment Instrument

Comparison	Pearson *r* Value
Part I & Part II	-.12
Part II & Part III	.20
Part I & Part III	.26

Next we analylzed the correlations between the individual parts of the Mid-Program Assessment instrument to identify the degree to which each part may be assessing unique competencies. As shown in Table 4, correlations were not significant and may be considered very weak (ranged from -.12 to .26). This suggests that each part of the Mid-Program Assessment instrument may be measuring different competencies. Given this finding, we limited our analysis to comparing only those scores derived from the Part II self-reports of the instrument with measures of observed performance.

Comparing Across the Participating Institutions

We next performed ANOVA and Tukey analyses (α=.95) to compare study measures across the different institutions. Study measures included the derived scores from the Mid-Program

Assessment instrument and performance measures generated from the coding scheme in Table 1. These measures were compared across institutions to determine the existence of significant differences. The results of these analyses were used to determine whether or not we could group our team data into a single population and therefore increase the statistical power of our results.

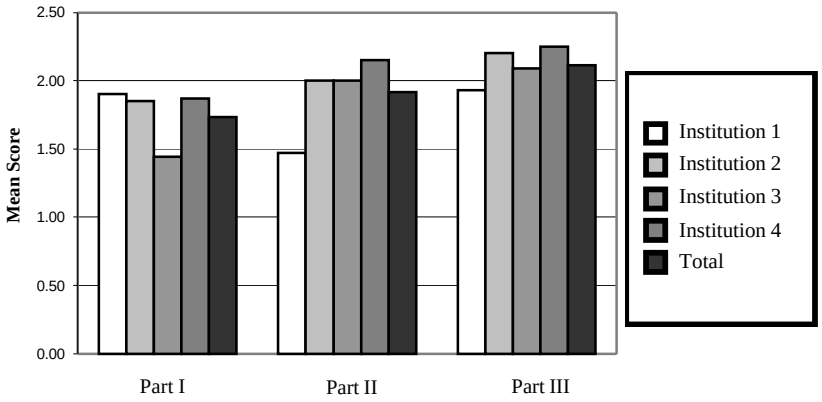


Figure 1. Mean Scores by Institutions for Part I, II, III of Mid-Program Assessment Instrument

Table 5. Differences in study measures by institution

Measure		Institution 1	Institution 2	Institution 3	Institution 4
Problem Definition	Mean	17.56	14.18	13.51	15.40
	Variance	7.40	18.67	1.06	8.10
Gathering Information	Mean	5.44	7.36	4.69	5.55
	Variance	10.22	5.39	8.59	3.82
Generating Ideas	Mean	1.42	0.69	1.45	1.28
	Variance	0.33	0.016	0.82	0.33
Modeling	Mean	4.77	6.69	9.79	9.13
	Variance	12.16	14.31	25.15	24.81
Feasibility[1]	Mean	1.02	0.53	2.21	0.78
	Variance	0.93	0.09	2.03	0.07
Evaluation	Mean	0	0.01	0	0
	Variance	0	2.89E-4	0	0
Decision	Mean	0	3.4E-3	0	0.12
	Variance	0	5.78E-05	0	0.03
Communication	Mean	0.38	1.10	1.00	1.04
	Variance	0.38	1.17	0.50	0.50
Transitions[2]	Mean	72.50	89.20	117	125
	Variance	30.45	6.30	24.25	26.81
Part II derived score	Mean	4.25	6.40	6.33	6.33
	Variance	0.92	1.30	2.33	4.67

Note[1]: Significantly different at p=.04
Note[2]: Significantly different at p=.015

As illustrated in Figure 1, there were only minor variations in derived scores across the participating institutions. A summary of performance measures by institution are provided in Table 5. As shown here, institutions were significantly different on only two, out of a possibe nine, performance measures: time spent in feasibility activities (p=.04) and number of transitions (p=.015). Given that groups spent a very small proportion of their total time in feasibility activities (under 2.5%) and that the feasibility code is only one of 8 possible design activity codes, this was not considered a crucial issue. However, the number of transitions represents an overall measure that has been found to be related to performance and academic standing in engineering (see Atman et al., 1999). As such, this is a more contentious issue when deciding whether or not to separate groups by institution. Upon closer examination, institutions that were most likely to be statistically different designed a testing procedure for a different hand tool (tree pruning shears rather than pet nail trimmers), had less variation across students' academic standing in the individual teams, and were more likely to have students of higher academic standing in the individual teams.

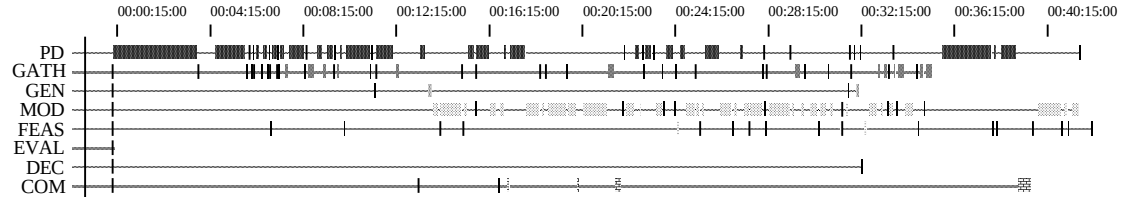
To probe deeper we repeated the analyses but separated groups by different hand tools rather than by institution. From this analysis it was found that groups that used the tree pruning shears had significantly more transitions (p=.019) and spent significantly more of their total time in decision activities (p=.016). Because only two of the nine performance measures differed across institutions, we decided to consolidate groups into a single population for this analysis. We plan to explore differences across problem type (hand tools) in future analyses.

Comparing Self-Reports to Observed Design Performance

The second level of analysis was a comparison of scores derived from the Mid-Program Assessment self-reports to coded measures of observed design performance. For this level, measures of design performance, as operationalized by the design activity codes in Table 1, were compared to scores from Part II of the instrument (see Table 3). Our guiding hypothesis was that teams that received credit for a particular TIDEE design element were more likely to spend time in design activities associated with that score. For example, teams that received a higher total score for Feasibility activities (TIDEE codes E3 and E5) would be expected to spend more time engaged in Feasibility activities. In addition, we expected to find a positive relationship between receiving a higher total score for teams that had more transitions and progressed farther into the latter stages of the design process.

To illustrate qualitatively, the timelines in Figure 2 represent the design activities of two design teams from the same institution (Team A and Team B). On the left side of each timeline are abbreviations for the design activity codes from Table 1. The tickmarks in the timelines represent time engaged in that coded activity. As shown in Figure 2, Team A received a higher score on the instrument and transitioned more across design activities. In comparison, Team B received a lower score on the instrument and transitioned less across design activities. As stated earlier, transitioning behavior has been shown to be a significant indicator of design success and greater engineering experience[25].

Team A: Part II Score = 9 (High), Number of Transitions = 159



Team B: Part II Score = 5 (Medium), Number o f Transitions = 86

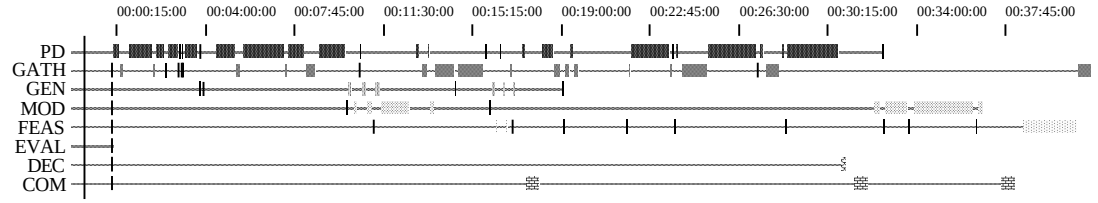


Figure 2. Timelines of coded design activity for two student teams from the same institution: Team A and Team B.

Table 6. Scores for Part II of the Mid-Program Assessment activity: Team A and Team B

TIDEE Code	TIDEE Decision Rule	Team A	Team B
AB1	Gathering tool information considered	1	
AB2	Time usage considered		
CD1	More than 5 ideas (customer expectations) listed	1	1
CD2	More than 3 solution ideas explained		1
CD3	One customer expectation selected	1	1
CD4	Selection of customer expectation is rationally justified	1	
E1	Relevant ideas for tests provided	1	1
E2	Detailed procedural steps for at least one test	1	
E3	Variability/replication of test results considered	1	1
E4	Quantification of test results provided	1	
E5	Criteria for passing test procedure provided	1	
	Total Derived Score	*9*	*5*

Exploring these trends further, the data in Table 6 supports our expectation that teams that received different scores from the Mid-Program Assessment instrument engaged in qualitatively and quantitatively different kinds of design activities. For example, Team A, unlike Team B, received credit for gathering information about the hand tool (e.g., the "problem") and various activities related to determining the feasibility and quality of their proposed design solution (e.g., quantified test results and criteria for passing the test). One interpretation suggested in Figure 2 and Table 6 is that because Team A engaged in a wider variety of design activities and transitioned more frequently between these activities they were more likely to receive a higher derived score from their self-reports.

To quantitatively compare derived scores from the team self-reports to measures of observed performance we utilized the mapping provided in Table 3 to conduct a series of ANOVA

analyses. The general format of each analysis was to compare the percent of total time engaged in a particular design activity with the mapped cumulative derived scores. For example, four derived scores (TIDEE codes CD1, CD2, CD3, and CD4) were mapped to Problem Definition activities. Therefore, the total score a team could receive ranged from a minimum of zero to a maximum of four. Based on our hypothesis, any score greater than zero should be associated with greater time spent in Problem Definition activities.

Table 7. Summary of comparison between derived scores and performance measures (N=18)

Design Activity	Derived Score (TIDEE Codes)	Derived Score (Cumulative)	No. of Teams that Received Credit	Percent of Time Spent in Design Activity	
				Mean	Std. Dev.
Problem Definition	CD1, CD2, CD3, CD4	0	0	0	0
		1	1	33.92	0
		2	12	47.14	13.5
		3	5	52.5	13.5
		4	0	0	0
Gathering Information	AB1, AB2	0	7	18.52	7.67
		1	10	18.24	8.66
		2	1	24.1	0
Generating Ideas	E1	0	2	3.12	0.94
		1	16	3.74	1.87
Modeling[1]	E2, E4	0	8	13.84	7.53
		1	4	29.54	9.83
		2	6	31.59	12.86
Feasibility[2]	E3, E5	0	12	2.54	2.01
		1	5	6.19	3.48
		2	1	2.76	0
Evaluation	E1	0	2	0	0
		1	16	0.01	0.03
Decision	E1	0	2	0.03	0.04
		1	16	0.14	0.35
Communication	E2	0	12	2.05	2.11
		1	6	4.26	2.48
Transitions	Total Score	0	0	0	0
		1	0	0	0
		2	0	0	0
		3	1	62	0
		4	2	128.5	16.26
		5	6	82.17	23.7
		6	4	116.25	31.66
		7	1	88	0
		8	2	102.5	14.85
		9	2	133.5	34.65
		10	0	0	0
		11	0	0	0

Note[1]: Significantly different at p=.01
Note[2]: Significantly different at p=.045

The results of the analyses for each of the mappings are summarized in Table 7. The percent of time spent in design activities generally increased as the amount of credit received from the scoring rubric increased. This relationship is illustrated in Table 7 for Problem Definition,

Modeling, Feasibility, Evaluation, Decision, and Communication activities. The number of transitions generally increased with the total cumulative score yet this was not a linear relationship. These results suggest that performance measures compare favorably with derived scores from the written self-reports. However, only two design activities were significantly and positively associated with receiving credit based on the scoring rubric: percent of time spent in Modeling activities (p=.01) and Feasibility activities (p=.045).

There were also some unexpected findings. Overall, teams spent a considerable amount of time engaged in Gathering Information activities. Nonetheless, teams that did not receive credit for Gathering Information activities spent 18.52% of their total design time in these activities. This was approximately equal to the amount of time for those who received at least some credit. Also, our findings illustrate that teams did not spend much time in Generating Ideas activities (about 3% of total design time). Like the Gathering Information measures, the results do not show a distinction between teams that received credit for these activities and the amount of time they engaged in these activities. Teams also spent a considerable amount of total time engaged in Modeling activities (ranged from 13.84% to 31.59%). More importantly, teams that received at least one credit for Modeling activities spent almost twice as much time as those that received no credit. Finally, teams that received the highest possible cumulative scores for Feasibility activities spent the same amount of time engaged in these activities as those who received no credit.

These findings present some interesting questions: why did teams that spent a considerable amount of time in particular design activities not receive any associated credit; why did teams that received credit spend similar amounts of time in associated design activities as those who received no credit? One interpretation is that teams were engaging in these activities but were not receiving credit due to missing information in their written self-reports. To pursue these questions we coded activities listed in the team design logs based on our design activity coding scheme (see Table 1). For the situation of Gathering Information this was a particularly useful analysis—the design log was the only component from the Part II self-reports that was utilized to determine whether or not teams received credit for Gathering Information activities (see Table 2). Examples of listed activities for which teams received credit include: “evaluated product, packaging, and instructions”, “read package, learn to use product”, “analysis of tool, find out what it is”, and “information gathering, examining tool”.

From this secondary analysis we found that teams that spent more time in Gathering Information activities were significantly more likely to identify these activities in their design logs (p=.004). We also found evidence of small errors in the scoring process; scorers were directed to not give credit for Gathering Information activities that included descriptive terms such as “testing”. For example, an activity such as “tested how easy the blade was to change and the grip” is clearly a Gathering Information activity. However, the existence of the word “test” conflicts with the established decision scoring rules[24]. As a result, this team did not receive the appropriate credit.

Overall, it was more likely that discrepancies associated with Gathering Information activities and receiving appropriate credit were a function of teams not documenting these activities in their self-reports. This finding was substantiated in the team videotapes. Many teams recorded their design activities either prior to or after completing the design task session; none of the

teams recorded their activities during the design task session. More specifically, at least five of the teams clearly stated that they wrote their design logs after completing the task and observations from nine of the teams indicate that they also wrote their design logs at the end of the task. Also, three of the teams were observed writing their design logs prior to selecting a customer expectation and developing an idea for a testing procedure. These design logs were essentially plans that may or may not have been followed. Explanations for why students did not capture Gathering Information activities in their self-reports may include: 1) teams did not *accurately* document their actual design activities, 2) teams may not have *recognized* that they were engaged in substantial Gathering Information activities that should be documented, 3) teams may have *erred in their recall* of their activities, 4) teams may not have perceived that these activities are *important or relevant* for describing their overall design processes. Each of these reasons has implications for design education. In particular, many researchers have studied information gathering behaviors in design problem solving. These findings provide evidence for a strong and positive relationship between the breadth and depth of relevant information designers gather with the quality of their final solutions and their level of experience[38]. Therefore, educating students about the importance of information gathering activities or increasing the effectiveness of these skills should be an important consideration in improving design education.

Our results also show that teams that spent time in Generating Ideas activities were significantly more likely to identify these activities in their design logs ($p=.003$). Nonetheless, teams received credit for Generating Ideas activities only if their chosen testing procedure was able to address the primary customary expectation (score E1 was based on "relevant ideas for test provided"). In other words, teams received credit for the quality of the ideas they generated, and perhaps not the quantity or the amount of time spent generating ideas. And finally, time spent in Modeling activities was significantly and positively correlated with the number of Modeling activities recorded in the team design logs ($p=.004$). There were only weak positive correlations for Feasibility (.387), Communication (.443), and number of transitions (.349).

Discussion

We began this study with a single focus: how do engineering students' written self-reports about design compare with observed design performance? The underlying goal was to assess the validity of a Mid-Program Assessment instrument for assessing design competency. The process of cross-validating the design component of this instrument (Part II) involved determining the boundaries of our study and our study population and then comparing coded measures of design performance to scores derived from students' written self-reports.

Overall, measures of observed performance compared favorably with scores derived from the Mid-Program Assessment instrument. In specific, there was a significant and positive relationship between the percent of time spent in Feasibility and Modeling activities and receiving credit from the scoring rubric for these activities. Other results suggest a positive, although weak, relationship between spending a greater percent of time in specific design activities and receiving the associated credit. However, we also found some significant disagreements. These were particularly the case for situations in which only a single scoring rubric element mapped to a design activity code. For example, there was evidence to suggest

that the assessment instrument was not capturing broad Gathering Information and Generating Ideas activities. Our findings indicate that many teams may have failed to adequately document their activities in their design logs. There was also evidence of small scoring errors.

Based on our findings, we collaborated with the TIDEE team to identify areas for improving the existing Mid-Program Assessment instrument as well as implications for improving design education. For the case of Gathering Information activities, the decision rules could be revised to include "testing" the hand tool specimen in order to gather information about the quality of the tool or how the tool is used. In addition, scoring elements could be added such that there would be more than one instance in the scoring rubric that maps to this activity. Similarly, the design log component of the instrument could be moved to the end of the activity to limit teams from writing design activity "goals" rather than documenting their actual design processes. There was also considerable evidence to suggest that students may need instruction in accurately documenting Gathering Information activities. This may involve educating students to recognize the role of information gathering in design success as well as activities to develop information gathering skills.

There are many strengths of this study. First, the measures for describing observed design performance build on a strong foundation of existing research. Secondly, the inter-rater reliabilities for both the coding of the observed performance and the scoring of the self-reports are very high which suggests that this study can be easily replicated. Third, the data in this study can be aggregated rather than analyzed by individual institutions which improves our ability to generalize our findings. There are also weaknesses. In particular, it can be difficult to have students in a team consistently talk aloud while performing a design task. Also, our measures for performance are limited to only verbal activities. Similarly, it can be difficult to get an accurate depiction of students' knowledge based solely on written assessments. In our case, we found that students may not always articulate and document what they know.

Conclusion

In summary, this paper is a report on the results of a study to cross-validate a mid-program instrument for assessing engineering student design competency. Our findings suggest that the relationship between students' self-reports and observed performance is not clear cut. In some instances there was considerable agreement across our study measures; in others, there were disagreements. We have also identified implications for improving the design of the assessment instrument as well as enhancing design education. In the future we plan to complement this research on the Mid-Program Assessment instrument with studies of design teamwork and communication.

Acknowledgements

This research was supported by funding from the National Science Foundation programs Course and Curriculum Development and Undergraduate Faculty Enhancement under grant number DUE 9455158. We would also like to acknowledge the contribution of the many people who participated in both data collection and data analysis: Denny Davis (PI), Steven Beyerlein, Robert Christianson, Joyce Cooper, Pat Daniels, Kenneth Gentili, Jana Littleton, Josh Martin,

Jeffrey McCauley, Larry Mckenzie, Sarintip Satitsatian, Robert Smith, and Michael Trevisan. Finally, we would like to thank the students who participated in this study.

References

1. American Society of Engineering Education (1994). *Engineering education for a changing world*. Engineering Deans Council and Corporate Roundtable of ASEE.
2. National Research Council (1995). *Engineering Education: Designing an Adaptive System*. Washington, DC: National Academy Press.
3. National Science Foundation (1995). *Restructuring engineering education: a focus on change*. Washington DC: National Science Foundation.
4. Accreditation Board for Engineering and Technology, (1998). *Engineering Criteria 2000: Criteria for accrediting programs in engineering in the United States* (2nd edition). Engineering Accreditation Commission, Accreditation Board for Engineering and Technology. http://www.abet.org
5. Campbell, S. & C. L. Colbeck (1998). Teaching and Assessing Engineering Design: A Review of the Research. *Proceedings of the Annual Conference of the American Society of Engineering Education*, Seattle, WA.
6. Shaeiwitz, J.A. (1996). Outcomes assessment in engineering education. *Journal of Engineering Education, 85* (3), pp. 239-246.
7. Terenzini, P.T., Cabrera, A.F., Colbeck, C.L., Parente, J.M. & S.A. Bjorklund (2001). Collaborative Learning vs. Lecture/Discussion: Students' Reported Learning Gains. *Journal of Engineering Education, January*, pp. 123-130.
8. Terenzini, P.T., Cabrera, A.F., Parente, J.M., & S.A. Bjorklund (1998). Preparing for ABET 2000: Assessment at the Classroom Level. *Proceedings of the Annual ASEE Conference*, Seattle, WA.
9. Byrd, J.S. & J.L. Hudgins (1995). Teaming in the design laboratory. *Journal of Engineering Education, 84* (4), pp. 335-341.
10. Van Duzer, E. & F. McMartin (2000). Building Better Teamwork Assessments: A Process for Improving the Validity and Sensitivity of Self/Peer Ratings. *Proceedings of the Annual Conference of the American Society of Engineering Education*, St. Louis, MO.
11. Dym, C. L. (1994). Teaching design to freshman: Style and content. *Journal of Engineering Education, 83* (4), pp. 303-310.
12. Miller, R.L. & B.M. Olds (1994). A model curriculum for a capstone course in multidisciplinary engineering design. *Journal of Engineering Education, 83* (4), pp. 311-316.
13. Rogers, G.M. & J. M. Williams (1999). Asynchronous Assessment: Using Electronic Portfolios to Assess Student Outcomes. *Proceedings of the Annual ASEE Conference*, Charlotte, NC.
14. Olds, B.M. & M.J. Pavelich (1996). A Portfolio-Based Assessment Program. *Proceedings of the Annual Conference of the American Society of Engineering Education*.
15. Davis, D.C., Gentili, K.L., Calkins, D.E. & M.S. Trevisan (1998). *Mid-Program Assessment of Team-Based Engineering Design: Concepts, Methods, and Materials.* Washington State University, Pullman, WA.
16. Davis, D.C., Gentili, K.L., Trevisan, M.S., Christianson, R.K. & J.F. McCauley (1999). Measuring Learning Outcomes for Engineering Design Education. *Proceedings of the Annual Conference of the American Society of Engineering Education*, Charlotte, NC.
17. Atman, C.J., Adams, R.S. & J. Turns (2000). Using multiple methods to evaluate a freshmen design course. *Proceedings of the Frontiers in Education Annual Conference*, Kansas City, KA.
18. Thompson, R. S. (2001). Reliability, Validity, and Bias in Peer Evaluations of Self-Directed Interdependent Work Teams. *Proceedings of the Annual Conference of the American Society of Engineering Education*, Albuquerque, NM.
19. Le Bold, W. & D. Budny (1995). How do students grade their learning. *Proceedings of the Frontiers in Education Annual Conference*, Atlanta, GA.
20. Baird, L. L. (1977). Using self-reports to predict student performance. *Research Monograph, 72* (4): pp. 33-39.

21. Cronbach, L. J. & F. C. Gleser (1957). *Psychological tests and personnel decisions*. Urbana, IL: University of Illinois Press.
22. Adams, R.S., Atman, C.J., Nakamura, R., Kalonji, G. & D. Denton (2002). Assessment of an International Freshmen Research and Design Experience: A Triangulation Study. *International Journal of Engineering Education, February* (in print).
23. Trevisan, M.S., Davis, D., Crain, R.W., Calkins, D.E. & K.L. Gentili (1998). Developing and Assessing Statewide Competencies for Engineering Design. *Journal of Engineering Education, April*, pp. 185-193.
24. URL: http://www.cea.wsu.edu/TIDEE.
25. Atman, C.J., Chimka, J.R., Bursic, K.M. & H. L. Nachtman (1999). A Comparison of Freshman and Senior Engineering Design Processes. *Design Studies,* Vol. 20, No., 2, pp. 131-152.
26. Adams, R.S. (2001). *Cognitive Processes in Iterative Design Behavior*. Dissertation in the College of Education, Seattle, University of Washington.
27. Ericsson, K. A., & Simon, H. A. (1993). Protocol analysis: Verbal reports as data. Cambridge, MA: The MIT Press.
28. Atman, C. J., & Bursic, K. M. (1998). Verbal protocol analysis as a method to document engineering student design processes. *Journal of Engineering Education, April,* pp. 121-132.
29. Cross, N., Christiaans, H., & K. Dorst (1996). *Analysing design activity*, John Wiley & Sons.
30. Mullins, C.A., Atman, C.J., & L. Shuman (1996). Freshmen engineer's strategies and performance when approaching design problems. *IEEE Transactions on Education*, October.
31. Atman, C.J. & K.M. Bursic (1996). Teaching engineering design: Can reading a textbook make a difference*? Research in Engineering Design*, 8, pp. 240-250.
32. Bucciarelli, L. L. (1996). *Designing engineers.* Cambridge: MIT Press.
33. Tang, J. C., & Leifer, L. J. (1991). An observational methodoloy for studying group design activity. *Research in Engineering Design, 2*, pp. 209-219.
34. Chi, M.T.H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences*, 6 (3), pp. 271-315.
35. Moore, P. L., & C. J. Atman (1995). Do freshmen design texts adequately define the engineering design process. *Proceedings of the Annual Conference of the American Society of Engineering Education*, Pittsburgh, PA.
36. Sanderson, P.M., Scott, J.J.P., Johnston, T., Mainzer, J, Watanabe, L.M., & J.M. James (1994). MacSHAPA and the enterprise of Exploratory Sequential Data Analysis (ESDA). *International Journal of Human-Computer Studies*, 41 (5), pp. 633-681.
37. Davis, D., Trevisan, M., L. McKenzie & S. Beyerlein (2001). Enhancing Scoring Reliability in Mid-Program Assessment of Design. *Proceedings of the Annual Conference of the American Society of Engineering Education*, Albuquerque, NM.
38. Bursic, K. M., & Atman, C. J. (1997). Information gathering: A critical step for quality in the design process. *Quality Management Journal, 4*(4).

Biographical Information

ROBIN ADAMS is a Research Scientist at the Center for Engineering Learning and Teaching in the College of Engineering at the University of Washington. She received her Ph.D. in Education from the University of Washington, her MS in Materials Science and Engineering from the University of Washington, and her BS in Mechanical Engineering from the California Polytechnic State University at San Luis Obispo. Her areas of research include cognitive processes in design behavior, iteration in design activity, designing assessment tools, and supporting a research-informed approach to engineering education. She was formerly the Local Evaluator at the University of Washington for the Engineering Coalition of Schools for Excellence in Education and Leadership (ECSEL), which was funded under the Engineering Education Coalitions Program of the National Science Foundation.

PIMPIDA PUNNAKANTA received her Bachelors and Masters degree in Industrial Engineering from University of Washington. During her Master degree, she served as a Research Assistant for the TIDEE (Transferable Integrated Design in Engineering Education) project.

CYNTHIA J. ATMAN is the Director of the Center for Engineering Learning and Teaching in the College of Engineering at the University of Washington. She also holds an academic appointment in Industrial Engineering. Dr. Atman received her PhD in Engineering and Public Policy from Carnegie Mellon University, her MS in Industrial and Systems Engineering from Ohio State University and her BS in Industrial Engineering from West Virginia University. Dr. Atman's research focuses on engineering education issues. She is an Associate Editor for the Journal of Engineering Education and was co-chair for the Frontiers in Education conference in 1997.

CRAIG LEWIS, Ed.D., is currently self-employed as an education and career counseling consultant in the state of Washington.