AC 2011-907: ESTABLISHING INTER-RATER AGREEMENT FOR TIDEE'S TEAMWORK AND PROFESSIONAL DEVELOPMENT ASSESSMENTS

Robert Gerlick, Pittsburg State University

Dr. Robert Gerlick is Assistant Professor of Mechanical Engineering Technology at Pittsburg State University.

Denny C. Davis, Washington State University

Dr. Davis is Professor of Bioengineering and Director of the Engineering Education Research Center at Washington State University. He has led numerous multidisciplinary research projects to enhance engineering education. He currently leads projects creating and testing assessments and curriculum materials for engineering design and professional skills, especially for use in capstone engineering design courses. He has been a Fellow of the American Society for Engineering Education since 2002.

Michael S. Trevisan, Washington State University

Dr. Michael S. Trevisan is Professor of Educational Psychology and Associate Dean for Research and External Funding in the College of Education. Dr. Trevisan is published widely in the fields of educational measurement and evaluation. In recent years, he has collaborated with Dr. Denny Davis to develop assessments for engineering education design courses.

Shane A. Brown, Washington State University

Shane Brown is an assistant professor in the Department of Civil and Environmental Engineering at Washington State University. His research focuses on conceptual understanding of engineering students and practitioners and conceptual change processes that lead to differences in understanding.

Establishing Inter-rater Agreement for TIDEE's Teamwork and Professional Development Assessments

ABSTRACT

Senior capstone design courses in engineering programs provide an opportunity to address important curricular objectives related to teamwork and professional development. In this course, students work within a team environment and are challenged with non-technical issues, such as communication, organization, self-directed learning, etc. By the end of their capstone experience it is hoped that students are prepared for the professional working environment. Capstone faculty, often with technical expertise in a specific branch of engineering, have expressed difficulty in teaching and assessing the types of knowledge, skills, and affective behaviors associated with these non-technical performance areas. When assessing teamwork, for example, the approach of "I know it when I see it" is not uncommon for an assessment process. Valid and reliable assessment instruments are needed for capstones which define expected performance criteria, and therefore offer guidance for teaching and learning. In addition to this formative use, summative assessments are also needed to document achievement of student growth with regards to these outcomes. To this end, collaborators from the Transferable Integrated Design Engineering Education consortium (TIDEE) have developed a suite of assessments for use in capstone courses, comprising four common performance areas: teamwork, professional development, design processes, and solution assets. For each of these areas of performance, multiple assessments have been developed and testing for validity and reliability has been ongoing. The purpose of this paper is to present results from a reliability study conducted with seven TIDEE assessments from the teamwork and professional development performance areas.

For each of the assessments tested, the degree of inter-rater reliability was determined, representing an estimate of the consistency of scoring between multiple raters. This type of reliability is significant for the TIDEE assessments as essay-type responses are elicited from students and, therefore, requires professional judgments by faculty to assess achievement. Each assessment was tested by having two faculty raters and two teaching assistant raters score a subset of student work with corresponding scoring rubrics. Percent agreement calculations and correlations were used to interpret the level of rater agreement. Interpretations of the results were made in light of the intended uses of each assessment: formative and/or summative. In general, the assessments were found to have scoring agreement of 85% to 100% within a one-point variation. Exact agreement ranged from a high of 60% to a low of 20%. Overall, the results indicated sufficient agreement for use with formative assessment (for enhancing teaching and learning). For summative use, five of the assessments should prove adequate in documenting student growth, including the Team Contract, Team Member Citizenship, Growth Planning, Growth Progress, and Professional Development assessments. The remaining two, Team Processes and Growth Achieved, may need to be revised to improve agreement. Suggestions for improvement include revisions to rubric descriptors for each level of performance, improved Frame-of-Reference rater training to decrease rater errors and increase accuracy, and, lastly, incorporation of Behavior-Observation-Training in the training protocol.

INTRODUCTION

Assessment is integral to effective teaching and learning as information gained through assessment allows instructors and students to gauge their progress and make necessary changes for continued improvement. In addition, assessment provides engineering programs information to gauge and document the achievement of stated learning outcomes, each of these being important components of ABET[1] requirements. The particular assessment instrument(s) developed and used for these purposes must therefore give users valid and reliable results on which decisions can be based.

To support these types of course-level and program assessment needs in engineering design, multiple assessment instruments related to design and lifelong learning outcomes have been developed by collaborators from the Transferable Integrated Design Engineering Education (TIDEE) consortium. Pilot testing has been conducted in recent years to study the validity and reliability related to TIDEE's instruments. In particular, validity studies have addressed the value of the assessments to users (instructors and students) from varied engineering disciplines while reliability studies have focused specifically on the level of inter-rater agreement (IRA) of scoring between multiple and diverse raters. The purpose of this paper is to present results from the IRA study conducted with the TIDEE assessment instruments. A total of seven of the fifteen TIDEE instruments are reported in this paper, including three related to teamwork and four related to professional development. The next sections give a general overview of the assessments included in this study followed by a description of the approach to testing for IRA, the methods used, and finally the results and discussion.

OVERVIEW OF TIDEE ASSESSMENTS

The TIDEE assessments are primarily intended for engineering capstone design courses in order to aid in the teaching and learning of professional skills and knowledge required of engineering graduates. In addition to this formative use, certain assessments were designed for summative purposes, allowing the achievement of outcomes to be measured for program and ABET accreditation documentation. Fifteen TIDEE assessments have been developed to address four critical performance areas in engineering design: teamwork, professional development, design processes, and solution assets. Table 1 presents a brief overview of these performance areas along with corresponding assessment instruments and the general performance criteria of each (adapted from Davis et al.[2]).

The TIDEE assessments typically incorporate multiple response methods including checklist, short answer, and essay. The Team Member Citizenship assessment, for example, asks students to assess themselves and their teammates with respect to important attributes of teamwork, as listed in Table 2. Students then assess the overall contributions of each member, including themselves, by assigning a relative value (in terms of a percentage contribution) to each member of the team. Following this, students then write essays describing a current strength and area for improvement of each team member. A similar approach is used in assessments within the professional development performance area. For the Growth Planning assessment, students first rate the importance and their current level of performance of a set of attributes identified for

Table 1. Summary of TIDEE's Capstone Design Course Assessment Instruments

Performance Area	Assessment Instruments*	General Performance Criteria
Teamwork: Team member contributions and team processes employed to support team productivity in design	1. Team Contract (F) 2. Team Member Citizenship (F) 3. Team Processes (F) 4. Teamwork Achieved (S)	Team member behaviors and team processes contribute to constructive relationships, joint achievements, individual contributions, and information management that synergistically yield high productivity.
Professional Development: Individual demonstration of improved knowledge, skills, and behaviors essential to engineering practice	5. Growth Planning (F) 6. Growth Progress (F) 7. Professional Practices (F) 8. Growth Achieved (S)	Individuals document professional development in technical, interpersonal, and individual attributes important to their personal and project needs, professional behaviors, and ways of a reflective practitioner.
Design Processes: Practices implemented that effectively and efficiently facilitate the production of valuable design project assets	9. Problem Scoping Processes (F) 10. Concept Generation Processes (F) 11. Solution Realization Processes (F) 12. Design Reflection (S)	Designers reflectively use design tools and information throughout problem scoping, concept generation, and solution realization activities to co-develop problem understanding and a responsive design solution.
Solution Assets: Design results that meet needs and deliver satisfaction and value to key project stakeholders	13. Defined Problem (F) 14. Selected Concept (F) 15. Proposed Solution (S)	Designers deliver and effectively defend solutions that satisfy stakeholder needs for functionality, financial benefit, implementation feasibility, and impacts on society.

* (F) indicates formative use; (S) indicates summative use

professional development, listed in Table 3. Next they write a short essay on one of these attributes which they feel is important to their professional growth and which needs to be developed further. The remaining TIDEE assessments within these two performance areas follow a similar format. Additional information and details for all of the TIDEE assessments can be found in references 3 and 4 and can also be found at www.tidee.org.

The assessments are administered to students through a web interface where members complete the assignment—either individually or as a team—and responses are then available to the instructor for review. A scoring rubric is used by the instructor to evaluate the responses and a summative score, along with instructor feedback, is returned to the student/team. Any or all of the assessments can be used throughout the duration of the course and the sequence and timing is determined by the instructor.

Table 2. Attributes for team member citizenship assessment

Category	Attribute/Ability
Team Relationships	1. Engages members with respect 2. Commits, encourages involvement 3. Resolves conflicts constructively
Joint Achievements	4. Helps establish shared goals 5. Follows plans to achieve team goals 6. Works synergistically with others
Member Contributions	7. Delegates/completes tasks, as needed 8. Performs competently to team standards 9. Enables development in self and others
Team Information	10. Strives for fully-informed members 11. Communicates well with stakeholders 12. Documents achievements well

Table 3. Attributes for professional development

Category	Attribute/Ability
Technical	1. **Analyzing information**: Applying methods/tools of analysis to understand and predict conditions 2. **Solving problems**: Formulating, selecting, and implementing actions for optimal outcomes 3. **Designing products**: Producing creative, practical products that bring value to varied stakeholders 4. **Researching questions**: Investigating, processing and interpreting information to answer important questions
Interpersonal	5. **Communicating**: Receiving, processing, sharing information in many forms to achieve desired impact 6. **Collaborating**: Working with a team to achieve collective and individual goals 7. **Relating inclusively**: Valuing and sustaining a supportive environment for all knowledge and perspectives 8. **Leading others**: Developing shared vision & plans; empowering to achieve individual & collective goals
Individual	9. **Practicing self-growth**: Planning, self-assessing, and achieving goals for personal development 10. **Being a high achiever**: Delivering consistently high quality work and results on time 11. **Adapting to change**: Being aware and responding proactively to social, global, and technological change 12. **Serving professionally**: Serving with integrity, responsibility and sensitivity to individual and societal norms

STUDY APPROACH

Reliability refers to the consistency of repeated measurements[5]. Operationally, this addresses to what extent test scores are consistent, dependable, and repeatable[6]. If an assessment is reliable, scores will be relatively stable over multiple administrations, given the trait measured doesn't change. The three basic methods for determining reliability include test-retest reliability, equivalent forms reliability, and internal consistency[7], and relate to the properties of the test. That is, the causes of measurement error are due to the quality of the items on the test, construction of the test, implementation process, etc. A fourth type of reliability, inter-rater agreement (IRA), refers to the level of agreement between multiple raters in scoring and is the method used for this study.

In objective tests, such as those with multiple-choice responses, correct responses are set and any errors in the scoring process are unintended. In subjective tests, where responses include essays, scoring is often dependent on the interpretations and standards of the rater (different raters may score the same response differently). IRA is significant for the TIDEE assessments since essay-type responses comprise a large part of the responses scored in each assessment. Additionally, as one of the primary objectives of the TIDEE assessments is transferability across multiple disciplines and diverse users, adequate IRA is an important quality. Consequences for mismeasuring student performance can lead to ineffective feedback, or possibly feedback which is counterproductive for students. For summative use, mismeasurement of performance can lead to inappropriate grading and, when used for ABET and program documentation, an incorrect basis for decisions.

Several methods are available for calculating IRA: percent agreement, Pearson's correlation, Cronbach's alpha, Cohen's Kappa, generalizability coefficient, intra-class correlation, and others. All of these methods have advantages and disadvantages and the choice of method depends on variables such as the number of test items, number of students/tests, range of scale (e.g., 2- or 7-pts.), type of scale (e.g., nominal, ordinal, interval, ratio), variance in scores, number of raters, etc.[8] For this study two measures are used to report and interpret the degree of IRA: percentage agreement (PA) and Pearson's product moment correlation coefficient (r). The choice of these two measures is based on the following factors, specific to this study and the TIDEE assessments: (1) the number of raters utilized for the study include two faculty raters and two teaching assistant (TA) raters, (2) the type of assessment scales for each associated scoring rubric are 5-point Likert scales (for all TIDEE assessments), (3) ease of interpretation, and (4) the purpose of the TIDEE assessments are primarily formative and, as such, higher-level statistical measures are not necessary as would be in large-scale, high-stakes assessment; that is, the measures chosen are sufficient as an indicator of the level of rater agreement which will be used to inform developers of the technical quality of the assessments and whether improvements are needed.

With the percent agreement method, the percentage of instances that multiple raters agree on a score, or within a score range, is given. This method is simple yet provides readers with results that are easily interpreted and meaningful, relative to more abstract coefficients of rater agreement. One problem with PA, though, is that the resulting value decreases with additional raters; three raters will likely have a lower agreement than two raters[8]. Also, the method does not

consider chance agreement. That is, with a 5-point scale, two raters scoring the same set of student work could have PA results based on chance alone of 20% exact agreement, 32% differing by one, and 48% differing by two; or, in terms of *within* a range of scores, 20% exact, 52% within 1 point of each other, and 100% chance that the two scores are within 2 points. To address the shortcomings of this method, this study will compare only two raters at a time and the results will be discussed in light of chance agreement statistics.

The Pearson correlation reflects the degree to which variables are related. Values for Pearson's range from +1 to -1, with +1 representing a perfect relationship, 0 no relationship, and -1 a perfectly negative relationship. This correlation provides a way to interpret the results in light of PA. For example, given a set of ratings which have low agreement, if the ratings do show a very high correlation, r (e.g., the two raters are consistently off by the same amount relative to one another), this would shed light on the nature of the error involved.

DATA COLLECTION METHODS

Participants and Data Sets

Pilot testing of the TIDEE Teamwork and Professional Development assessments was conducted throughout the 2008-09 and 2009-10 academic years with collaborators from six universities. Participating institutions varied in size and background and included disciplines from mechanical, civil, bioengineering, general engineering, and others. The faculty from these institutions selected various TIDEE assessments to implement during their capstone course—typically consisting of two assessments from the Teamwork area and two from the Professional Development area—through the web-based TIDEE system. Student responses and resulting faculty scoring and feedback were then made available to TIDEE researchers. From this data, a sample of student work was collected and used to conduct the IRA study.

The IRA study was conducted during the Spring of 2009 with the three formative Teamwork assessments—Team Member Citizenship, Team Processes, and Team Contract—and all four of the Professional Development assessments (see Table 1). For each assessment, a set of student work was selected from the available data and scored by four raters—two faculty and two TAs. Criteria for faculty participation included being in an engineering discipline and having experience teaching capstone design courses. A total of fourteen faculty participants were recruited from nine different engineering colleges throughout the US. This allowed faculty raters to vary for each of the assessments. For the TA raters, two were used for all seven assessments tested. One student was an engineering doctoral candidate in Chemical Engineering while the other was a senior in Bioengineering who had previously taken the capstone design course and had experience with the TIDEE assessments.

Student work selected for use in the study was actual work submitted by students from the six collaborating schools. For each assessment, twenty pieces of student work were selected for testing with an additional two pieces selected for rater training. The selection process for student work included, first, reviewing all work available, discarding any work with no student responses

submitted (only those with at least partial completion were considered), and then randomly choosing twenty pieces from those remaining.

Study Procedure

For each of the seven assessments, two faculty raters were randomly assigned to one assessment for scoring. Each rater was contacted by email and a convenient date and time coordinated for the study, on an individual basis. A packet was mailed to each rater prior to the study. Packets contained a brief overview of the purpose of the study, rater instructions, two pieces of student work for rater training, twenty pieces of student work for the study, a scoring sheet (with rubric) for each of the twenty works, and a return envelope for mailing scoring sheets back. At the scheduled date and time, each rater was contacted by telephone and given training on the assessment and scoring process; all raters received individual training and conducted the scoring independent of other raters. Rater training consisted of two types of training: rater-error training (RET) and frame-of-reference (FOR) training.

In general, the goal of RET is to improve rater accuracy by reducing errors, or rater biases. Common types of errors include halo, leniency/severity, and central tendency, among others. RET focuses on familiarizing raters with these types of errors through exemplar works and then encouraging raters to avoid them[9]. For this study, RET was accomplished by first giving a brief description of these typical errors and then instructing raters on appropriate scoring procedures (discussed below). RET represented a small portion of the total training protocol.

FOR training was used more extensively in this study. This type of training involves orienting raters with the norms and standards expected for the particular behavior being evaluated. That is, in addition to instructing raters on performance dimensions/behaviors being assessed, exemplars of critical incidents are also presented which give raters an understanding, or, frame-of-reference, of the performance in the specific context of the organization (in this case, capstone design courses). The FOR training process can be described as the following sequence of activities: defining the performance dimensions, providing multiple incidents of behavior for each dimension and descriptions of the levels of performance for each (through several vignettes; each at different levels of performance), rater practice in evaluating performance followed by discussions on the discrepancies between participant ratings and true ratings[10]. Some other training approaches also used in performance appraisal include behavior observation training (BOT) and performance dimension training (PDT)[9]. However, FOR training was selected for this study, as it has been reported to result in higher inter-rater agreement and accuracy[11, 12].

The rater training protocol used in this study consisted of the following activities for each rater participant:

1. Review the purpose and significance of the IR agreement study.
2. Closely review and discuss the assessment to be scored and its associated scoring rubric.
3. Give instructions and tips for scoring (addressing RET; listed below).
4. Rater reads through one piece of student work (a calibration work, which is in addition to the twenty pieces used for IRA calculations) and scores with rubric.

5. Review rater's scores and justifications with those of trainer; discuss differences and reach consensus.
6. Repeat 4 and 5 with a second calibration work.

Instructions and tips for scoring, including those related to RET, were given in the training session and consisted of the following:

- Circle one, and only one, descriptor for each criterion in the rubric.
- Take 4 to 5 minutes to score each one of the twenty pieces of work—no more or less.
- Score all work in a single session.
- Score each criterion individually, each based on its own merit.
- Rate absence of material according to criterion descriptors.
- Do not rate students against each other.
- Score work according to the rubric descriptor; don't let other factors affect scores (e.g., writing quality, perception of student's understanding, quantity of writing).

Following this rater training, each rater was instructed to score the work within the next two days. Raters then returned the twenty score sheets in the return envelope provided. The two TAs were given the same training as the faculty raters for each for the seven assessments and their scoring of work was completed in the same sequence as the faculty raters.

RESULTS AND DISCUSSION

For each of the seven assessments tested, a set of scores was obtained from each of the four raters. Each of these sets included twenty individual ratings: one per student for each of the twenty pieces. Calculations of PA were made between the two faculty scores and then between the two TA scores for both exact agreement as well as for agreement within one point. These results are presented in Figure 1 below in graphical and tabular form.

Results presented in Figure 1 beg several observations. To begin with, five of the seven assessments show faculty raters (labeled *Instr* in Figure 1), with exact agreement above 20%, the level of agreement based on chance alone. Interestingly, exact agreement levels by the TAs were well above 20% for all of the assessments. Additionally, for all but one assessment, Team Contract, the levels of exact agreement for the TAs were higher or equal to those of the instructors. In terms of agreement within one point, both instructors and TAs showed much higher agreement than what would be expected based on chance alone, which is found to be 52% for a 5-point Likert scale. Lastly, no rater-pair scores differed by more than two points for any of the assessments.

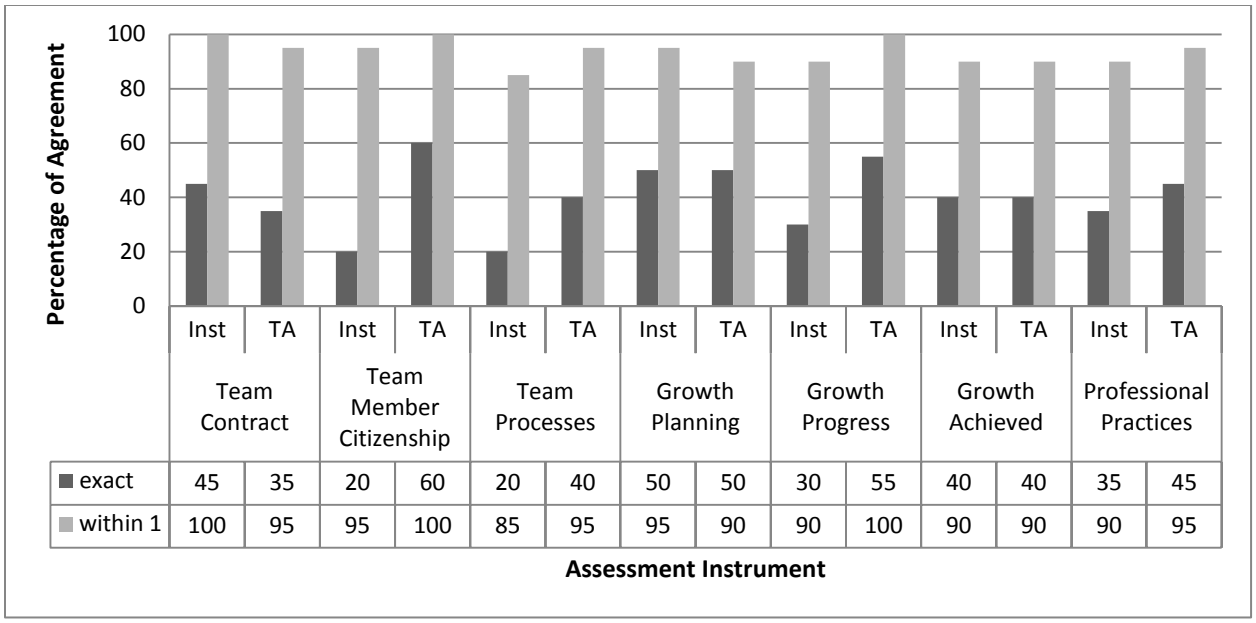

Figure 1. Results of percent agreement between raters

One challenge to interpreting the results of this study is that no standard exists that signifies the acceptable range of percent agreement levels. This may be due, in part, to the various contexts in which assessments are used as well as the various purposes for assessment; for example, large-scale, high-stakes summative assessment versus classroom-level formative assessment. For the intended formative purposes of these TIDEE assessments, the 85 to 100% agreement found for the *within one point* range would likely be quite adequate. The intent of the assessments is to support the types of instruction/learning decisions and goals of instructors and students with regard to teamwork and professional development in the context of engineering design. Since it is highly likely (a 85 to 100% chance) that any given rater will score work within one point of another rater, students (irrespective of rater) are very likely to get similar and consistent feedback on their performance. Based on this observation, the level of agreement found between raters should be considered acceptable for each of the seven assessments tested in this study.

In addition to this evaluation of the assessments, the results can also shed light on the degree of transferability of the assessments, an important goal of TIDEE collaborators. Faculty raters were from very diverse backgrounds and undoubtedly had many differences that could have influenced their ratings. For example, differences between any two raters throughout this study likely included many of the following: engineering discipline, levels of involvement in capstone design courses (and therefore different knowledge of the related outcomes), years of experience in industry and academia, perceived importance of the capstone design experience, pre-existing knowledge of teamwork/professional development, pre-existing standards of student performance, attitude toward assessment, attitude toward students, and probably many others. In light of these differences, the degree of rater agreement found shows transferability among different raters. This conclusion can be extended to TA raters also as their agreement was as high or often higher than those of instructors, an important finding as TAs are anticipated to be involved in the scoring of assessment work.

For continued improvements to the TIDEE assessments, sources of rater differences need to be identified. To aid in this analysis, values of Pearson's correlation coefficient were found and are presented in Figure 2. In contrast to the higher levels of rater agreement of TAs over instructors mentioned above, the correlation of ratings between instructors was higher or equal to those of TAs for five of the six assessments. Therefore, while TAs more often agreed with one another on ratings, instructors were more consistent in relative rankings of performance. A possible explanation for this is that TAs, with little engineering experience to draw on or experience in observing student teams in a capstone environment, may have referred to the scoring rubric more closely and more often. Conversely, instructors may not have applied the rubric as strictly as TAs. Given their greater experience with capstone teams and with grading in general, instructors may have understood and adequately applied the relative rubric graduations, but with their own performance standards factoring in as well. With Team Member Citizenship, for instance, TA raters scored exactly 60% of the time while instructors were at 20%. In terms of rankings, though, instructor's scores were more correlated than TA's by .57 to .45.

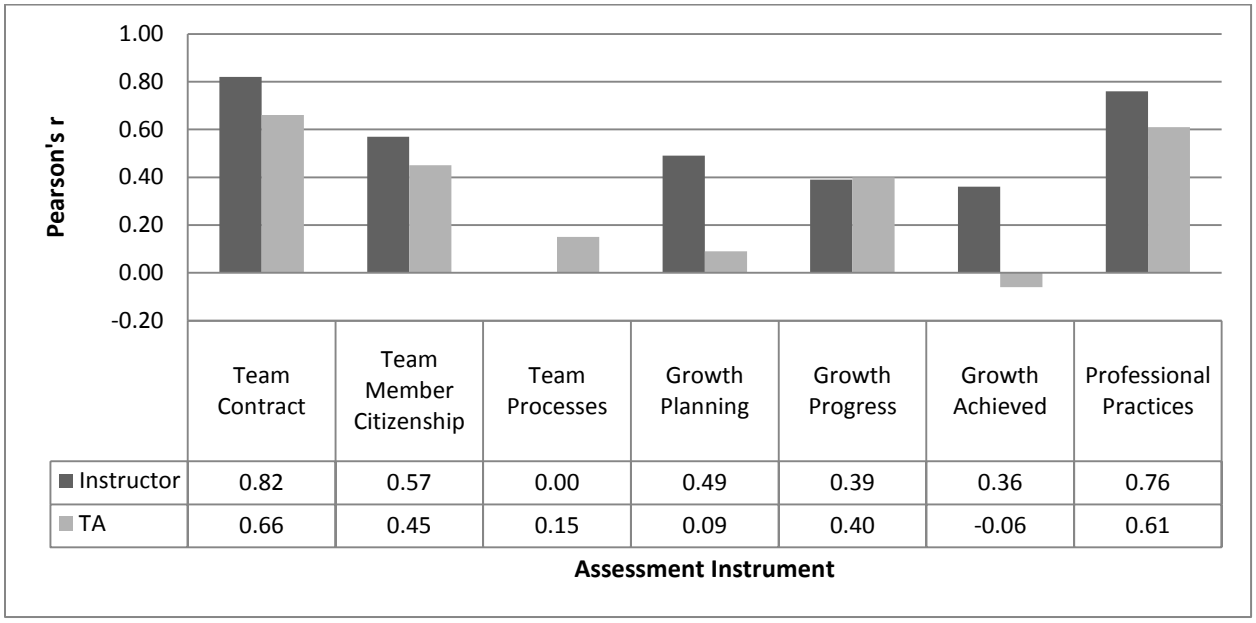

Figure 2. Results of Pearson's correlation between raters

Of the seven assessments tested, the Team Processes was found to have the lowest correlations: 0.00 for instructors and 0.15 for TAs. In terms of PA, instructors agreed exactly 20% and 85% within one point while TA's were 40% exact and 90% within one. For the Growth Achieved assessment, the TAs also had a low correlation, -0.06, but with sufficiently high agreement for both sets of raters. As mentioned above, for each of these assessments the higher levels of agreement within one point should prove adequate for the purpose of achieving teaching and learning goals. The rubrics will guide various raters toward similar interpretations of student work and, subsequently, provide similar types of feedback to students as well as consistent feedback on teaching practices. The low correlations, though, may indicate that the rubric

descriptors need further specificity, giving raters more detailed instructions on ratings. This approach, though, would decrease generalizability, representing a common tradeoff in assessment/rubric design.

CONCLUSIONS AND RECOMMENDATIONS

The IRA results presented in Figure 1 along with the corresponding correlations in Figure 2 indicate areas of strength and possibilities for improvements. In particular, all assessments show levels of rater agreement likely to be sufficient for the purposes of formative assessments in capstone design courses, as shown by the high levels of agreement within one point. The Team Processes and Team Member Citizenship assessments show the lowest levels of exact agreement between instructor raters, although, given the high agreement within one point and the formative use of the assessments, this result should be considered acceptable. The Growth Achieved assessment, on the other hand, which is intended more so for summative use, shows a correlation of -.06 for TA raters. And, as mentioned, since TAs are likely to be involved in the ratings of student work, decreasing this variability in ratings for this assessment would be beneficial.

To improve IRA, qualitative analysis of each rater's actual scoring approach, obtained through post-scoring interviews, would certainly shed greater light on possible sources of errors. Errors may be related to the rubric design (e.g., exclusiveness of anchors), to the varied backgrounds of raters, or to rater training. Without any certain indication of the source of errors, and hence clear direction to take in improving agreement, addressing rater training in any case would undoubtedly lead to improvements in rater agreement, as reported by several researchers in this field[9, 10, 11, 12, 13, 14]. The training provided to raters during this study was a scaled down version of Bernardin and Buckley's frame-of-reference (FOR) training. For example, Bernardin and Buckley suggest that three vignettes be used to provide a frame-of-reference for three distinct levels of performance: high, medium, and low. In this study, only two were used, drawn from actual student work and thus not necessarily the optimal exemplars for training purposes. It is also suggested by Bernardin and Buckley that, prior to reviewing the three vignettes during training, raters develop (through group discussion) a set of performance behaviors for the target as well as the corresponding levels of performance for each. This would require raters to reflect on the performance more deeply, which would presumably lead to greater awareness of the behaviors during scoring. During this study, raters did not go through this activity. After reviewing the assessment, the rubric was reviewed only briefly (not every detail was discussed). With the training process in this study typically lasting between twenty and thirty minutes, a practical balance must be reached between breadth and depth as this timeframe was probably on the longer side of what raters considered acceptable.

Therefore, one recommendation to improve rater agreement would be to design a training protocol which provides trainees with a very clear understanding of performance expectations, shown through tailor-made critical incident vignettes. A second recommendation would be to include elements of behavior observation training (BOT) into the training protocol as well. This type of training gives raters skills and experience in observing behaviors effectively. This is suggested since, without knowledge of what to observe, raters are said to rely on their own experiences to decide what to observe[11]. This could allow deviation to occur in methods of

observation, and, therefore, deviation in scores. Additionally, BOT may be particularly beneficial to formative assessment as raters are encouraged to notice particular behaviors, which could then lead to more pointed feedback.

Rubric anchors for the Growth Achieved assessment could also be revised so that the descriptions of each level of performance are more robust, leading to higher correlations among scores. Although this may result in less flexibility in scoring—a priority of instructors—greater consistency is needed as this particular assessment is intended for summative purposes. The low levels of exact agreement, coupled with relatively low correlations of the Team Processes assessment, warrant similar adjustments.

Lastly, a recommendation for further study relates to the variables affecting faculty scoring agreement. These potential variables may include (as listed above): engineering discipline, levels of involvement in capstone design courses, years of experience in industry and academia, perceived importance of the capstone design experience, knowledge of teamwork and professional development concepts, standards of student performance, attitude toward assessment, and attitude toward students. With transferability among diverse faculty and disciplines representing an important goal for the TIDEE assessments, a clearer understanding of the causal relationships associated with variables such as these may suggest areas for improvement in rater training, administration practices, and/or assessment content.

ACKNOWLEDGEMENTS

This work has been supported by a grant from the National Science Foundation (NSF), Division of Undergraduate Education: *DUE-0919248*. The authors express sincere gratitude to NSF for this financial support.

REFERENCES

1. ABET. (2009). *Criteria for accrediting engineering programs*. Retrieved November 1, 2009, from http://www.abet.org/forms.shtml
2. Davis, D., Beyerlein, S., Thompson, P., Harrison, O., & Trevisan, M. (2009). Assessments for Capstone Engineering Design [Electronic Version]. Retrieved July 1, 2010 from www.tidee.org.
3. Davis, D., Trevisan, M., Davis, H., Gerlick, R., McCormack, J., Beyerlein, S., Thompson, P., Howe, S., Leiffer, P., Brackin, P., & Khan, J. (2010). Assessing professional skill development in capstone design courses. *Proceedings of the Capstone Conference*, Boulder, CO.
4. Davis, D., Trevisan, M., Gerlick, R., Davis, H., McCormack, J., Beyerlein, S., Thompson, P., Howe, S., Leiffer, P., & Brackin, P. (2010). Assessing team member citizenship in capstone engineering design courses. *International Journal of Engineering Education, 26*(4), 1-13.
5. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
6. Drummond, R., & Jones, D. D. (2006). *Assessment Procedures for Counselors and Helping Professionals* (6 ed.). Upper Saddle River, NJ: Pearson Education.
7. Kubiszyn, T., & Borich, G. (2003). *Education Testing and Measurement: Classroom Application and Practice* (7 ed.): John Wiley and Sons, Inc.

8. Abedi, j., Baker, E. L., & Herl, H. (1995). *Comparing Reliability Indices Obtained by Different Approaches for Performance Assessments*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
9. Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205.
10. Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6*(2), 205-212.
11. Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86*(2), 255-264.
12. Stamoulis, D. T., & Hauenstein, N. M. A. (1993). Rater training and rating accuracy: Training for dimensional accuracy versus training for ratee differentiation. *Journal of Applied Psychology, 78*(6), 994-1003.
13. Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. *Journal of Applied Psychology, 64*(4), 410-421.
14. McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology, 69*(1), 147-156.