Work in Progress - Improving Interrater Agreement Used to Measure Learning Outcomes

Heling Shi, Daniel Ferguson, Jonathan Beagley, Margaret Huyck
Illinois Institute of Technology, shiheli@iit.edu, dferguson@iit.edu, beagjon@iit.edu, mhhuyck@sbcglobal.net

Abstract - The Inter-professional Project Program (IPRO) at IIT hosts a one-day evaluation event (IPRO Day) every semester, during which 70-120 judges (~50% new each semester) judge 35-45 teams averaging 12 students each. The judges, in groups of three to five, rate ten criteria per team on a 1-5 scale for either group oral presentations or exhibits. While the statistics of measuring interrater agreement are well documented it is not completely clear from previous research how best to improve such agreement except for assertions that rater training is beneficial. Over the past two years, we have instituted several interventions to improve inter-rater agreement on IPRO Day: using experienced judges to coordinate each judging group, rewriting judging forms to match the learning objectives, developing rubrics to guide judging, and conducting orientation sessions with the judges. This study evaluates the effectiveness of these interventions by comparing judging group agreement results from semesters prior to the change with the semesters following the change. These results having broad applicability to all judging review panels.

Index Terms - Evaluation techniques, interrater agreement, interventions.

INTRODUCTION

The IPRO program at IIT, since its inception in 1995, has become a signature program of the school for it not only benefits participating students but also project sponsors and prospective employers of IIT graduates. The program is designed to engage students in a multi-discipline, teamwork based learning environment to solve a real-world problem proposed either by industry sponsors and/or IIT faculty. Additionally, the program helps students develop and apply their teamwork, project management, communication and ethical awareness skills. At the end of the semester, as a part of the IPRO learning outcomes (LOs) assessment system, students participate in IPRO Day to exhibit and present their project course work to assigned judges.

IPRO Day is designed to measure team achievements in all of the learning objectives. Each semester, all IPRO teams are required to prepare both a 20-minute oral presentation and an exhibit staffed for six hours which explains they have accomplished during the semester. Independent judges, who are drawn from faculty, alumni, professionals and graduate students, are recruited, trained and given judging sheets to use to evaluate the learning objective achievements of the teams. Each judging sheet contains ten criteria and nine of which are identical between presentations and exhibits. These criteria refer specifically to the learning objectives. There are three criteria focused on project management, one each on communications, ethical awareness and teamwork and three on project results.

HISTORY OF JUDGING INTERVENTIONS ON IPRO DAY

A number of interventions have been instituted in attempt to improve the inter-rater agreement of IPRO Day judging. Given that about 50% of the judges each semester have judged previous IPRO Day events, chief judges are selected from this pool to each facilitate a judging group by arranging the group to confer on the rating to reduce inter-rater variances. All judges are also asked to attend either the one-hour orientation hosted at the start of the day and/or use the online judges training system. A set of rubrics for each criterion, identifying how to classify teams on the 5-point scale, was created and provided to the judges. Each team is judged by 3-5 judges for their presentation and 3-5 judges for their exhibit, a total of 6-10 judges overall.

DESCRIPTION AND EVALUATION OF IRA

Interrater agreement (IRA), along with interrater reliability (IRR), is a powerful measure of the consistency and quality of rating/judging events, particularly during the evaluation of learning outcomes in an educational setting. Both IRA and IRR signifies the extent to which two or individual raters/judges view the same phenomenon or object and agree on the same opinion or score, and are crucial when humans are utilized as the measuring instrument at such events, like that on IIT’s IPRO Day [1]-[3]. While IRR focuses on the relative consistency in the scores provided, IRA focuses on the absolute consensus in the scores and whether if the scores by judges are interchangeable in terms of their absolute value. For a multi-target assessment modeled judging event like IPRO Day, estimates of IRA are required to provide sufficient conclusions of the quality the context rating/judging event [4]. A system with low or unstable IRA would considerably undermine the validity of its program outcomes thus suggest inconsistency in the scores and raise question of the fairness of the scoring system.

The level of interrater agreement is typically assessed via the $r_{WG}$ method, which does not take in consideration of
the varying number of judges or the varying point-scale per item. Therefore, the aWG index, a method that suggests IRA is also influenced by the number of judges and number of points in the rating scale, was used to interpret the IRA of the data collected. The direct computing equation derived by Brown and Hauenstein [5] is used:

\[ a_{WG} = 1 - 2S^2/[(H+L)X^2-(H*L)]*[K/(K-J)], \]  

(1)

where \( a_{WG} \) is the IRA by question/item, \( S^2 \) is the mean of the observed variances on the \( J \) items being judged, \( X \) is the observed mean, \( H \) is the maximum possible value of the scale, \( L \) is the minimum possible value of the scale, and \( K \) is the number of judges.

**METHODOLOGY**

The data analysis reported follows the procedure designed by Brown and Hauenstein to measure inter-rater agreement. The \( a_{WG} \) algorithm is applied to IPRO day judging data for three semesters, spring 2007 to spring 2008, as shown. The results are then related to the history of judging interventions attempted during each semester to evaluate their corresponding effectiveness.

Vargo et al’s [6] experiments on E-Learning learning object evaluation also support the interventions being evaluated and suggest that modifications be made prior to and during the judging process, such as training sessions for the raters, assuring a minimum number of raters with matched expertise, prioritizing post-scoring discussions before the submission of scores, and revising grading rubrics based on user recommendations; all improve IRA measurements.

**DATA ANALYSIS**

Using (1), the IRA by question/item was calculated for the IPRO day data from spring 07, fall 07 and spring 08. Uniform distribution of the data is assumed during this calculation. This analysis of IRA by item shows a different level of agreement as depicted in Table 2 where the level of agreement varies by question and a comparison between fall 07 and spring 07 shows an increase in variability of judges ratings, most likely due to new rating categories introduced in fall 07 as well as a change from a 10-point to a 5-point rating scale. Whether also providing rubrics had a positive or negative effect on IRA cannot be determined as the use of rubrics was introduced simultaneously with the change in rating scales.

**RESULTS**

Table 2 shows a significant decrease in IRA in the fall 07 data comparing both exhibits and presentations from that of spring 07. Matching the numbers with the timeline of interventions during the specified period, (see row 1, Table 1), we presume that the interventions carried out were effective (rewritten LOs and seeking conferment on ratings within judging groups). As for the decrease in IRA in the fall 07 data, newly introduced IPRO Day criteria and rubrics and their unfamiliarity to the judges may be one cause for increases in variability and therefore decreases in IRA.

**CONCLUSIONS/FUTURE WORK**

IRA in the range of 0.71 to 0.90 is deemed high and shows strong agreement and the interventions prior to fall 07 seem to have produced a rating process with very high IRA. The latest changes, if permanent, will have decreased IRA, opposite of what was intended by the changes. The data from spring 2008 show continued degradation in IRA, and therefore, the 5-point scale would be discarded and the judging process returned to its prior format. However, the changes to judging questions were also made to better measure and reflect the real-world accomplishments of the teams. It may be that the judges have a learning curve problem and that the IRA will rise again as they become more familiar with the new criteria.

**REFERENCES**


