

# Towards effective group work assessment: even what you don't see can bias you

Gahgene Gweon<sup>1</sup> · Soojin Jun<sup>2</sup> · Susan Finger<sup>3</sup> · Carolyn Penstein Rosé<sup>4</sup>

Accepted: 15 September 2015 / Published online: 23 September 2015  
© Springer Science+Business Media Dordrecht 2015

**Abstract** In project-based learning (PBL) courses, which are common in design and technology education, instructors regard both the process and the final product to be important. However, conducting an accurate assessment for process feedback is not an easy task because instructors of PBL courses often have to make judgments based on a limited view of group work. In this paper, we provide explanations about how in practice instructors actually exhibit cognitive biases and judgments made using incomplete information in the context of an engineering design education classroom. More specifically, we hypothesize that instructors would be susceptible to human errors that are well known in social psychology, the halo effect and the fundamental attribution error, because they have a limited view of group work when they facilitate distributed and remote groups. Through this study, we present two main contributions, namely (1) insights based on classroom data about limitations of current instructor assessment practices, (2) an illustration of using principles from social psychology as a lens for exploring important design questions for designing tools that monitor support oversight of group work. In addition to the study, we

---

✉ Soojin Jun  
soojinjun@yonsei.ac.kr

Gahgene Gweon  
ggweon@kaist.ac.kr

Susan Finger  
sfinger@cmu.edu

Carolyn Penstein Rosé  
cprose@cs.cmu.edu

<sup>1</sup> Knowledge Service Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro, Yuseong-gu, Daejeon, Republic of Korea

<sup>2</sup> Graduate School of Communication and Arts, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 120-749, Korea

<sup>3</sup> Civil and Environmental Engineering, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA

<sup>4</sup> LTI, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA

illustrate how the findings from our classroom study can be used for effective group assessments.

**Keywords** Instructor assessment · Halo effect · Fundamental attribution error · Project-based learning · Group work

## Introduction

Project-based learning (PBL), especially in courses where students work in groups on real world problems in collaboration with industry sponsors, is commonly believed by educators and administrators alike to have great value (Dutson et al. 1997; Rohde et al. 2007). One critical aspect, which impacts the educational value of such a course, is good feedback from instructors that allows students to learn a model process for carrying out projects (Bober et al. 1998; Nicol and Macfarlane-Dick 2006; Price et al. 2011; Gómez Puente et al. 2013). In particular, in PBL courses that are common in design and technology education, instructors regard both the process and the final product to be important. Thus, research on developing appropriate assessment tools and processes to aid instructors with the process assessment has been carried out. For example, Project E-scape examined methods for fostering learners' innovative performance and teamwork. As an outcome of the project, real-time design e-portfolios are created to capture the learners' design activity and to assess the final products (Kimbell 2007).

However, conducting an accurate assessment for feedback is not an easy task because instructors of PBL courses often have to make judgments based on a limited view of group work. Therefore, instructors may make an assumption by inferring students who are strong in a certain observed skill are also strong in an unobserved area. For example, literature shows that gregarious students make a better impression on instructors and are regarded as contributing more (Gopinath 1999). Yet assessments based on such impressions are susceptible to cognitive biases. This potential problem faced by the instructors, namely the problem of cognitive biases that influence assessment under conditions of preconceived and limited information, has also been explored by psychologists (Cooper 1981; Harvey et al. 1981; Becker and Cardy 1986; Burger 1991; Balzer and Sulsky 1992). More specifically, the types of cognitive biases that are applicable in our situation are perception biases that are measured from people with expertise (instructors) towards people in who are being assessed (students). In this paper, we explore two types of cognitive biases that address this issue: the halo effect (Thorndike 1920) and the fundamental attribution error (Ross 1977). The halo effect, is a cognitive bias leading people to perceive another person's traits in a way that is consistent with previous impressions of other traits (Thorndike 1920). The fundamental attribution error is the frequent tendency for people to underestimate the effect of context and to overestimate personality-based influences when explaining behaviors observed in others (Ross 1977).

In particular, we would like to examine whether these two cognitive biases are observed when instructors in a PBL setting make assessments and if so, what the extent of these biases are. To explore these research questions, we designed a classroom study that compares the difference in perspectives among instructors and observers, who can view the students' group work environment first hand. Applying findings from social psychology to support assessment practices in a PBL setting has been also reported by Johari and

Bradshaw (2008). They used three motivation theories, namely attribution theory, self-efficacy theory, and self-determination theory, to develop guidelines for assessing a PBL internship program. For example, they showed that mentors of the internship program could provide good feedback to students by reminding them of past successful experiences based on self-efficacy theory. However, Johari and Bradshaw did not explore cognitive biases on assessment practices or illuminate struggles faced by project-based learning course instructors in a classroom setting. Since the impact of cognitive biases is a relatively unexplored area in a PBL context, our work could motivate a broader investigation of assessment practices. More specifically, the context of our classroom study is in a graduate level capstone engineering design course that offers students the opportunity to integrate the knowledge and technology they have acquired in lower level design or technology courses.

In addition to exploring the impact of cognitive biases in graduate level project course, results from our classroom study would be valuable in the design of assessment tools. To address assessment problems that instructors face, there has been great and increasing interest in developing automatic visualization tools in the form of dashboards to aid facilitators of group work assessment (Gipps 2005; Jochems and Kreijns 2006; Wong et al. 2007; Ren et al. 2008; Phielix et al. 2010). It is exciting to see the advancement of educational technology in terms of technical aspects of such assessment tools. However, it is equally crucial for the developers to be aware of cognitive biases that prevent instructors from making an accurate assessment in the classrooms so that the developers can provide appropriate assessment tools that capture such bias. Therefore, results from our classroom study can be used to provide motivation for researchers to then seek to build tools that address a root cause of difficulty, in this case, to help facilitators to be less susceptible to cognitive biases.

In the rest of the paper, we start by presenting the details of two hypotheses regarding the halo effect and the fundamental attribution error (“[Research motivation and hypotheses](#)” section). Next, details on how the classroom study was conducted are presented in “[Method for the classroom study](#)” section. A classroom environment is noisier than an experimental environment in that many more variables are in play. However, in a simulated group, members do not have a vested interest in maintaining their long term relationship. Therefore, only conducting a classroom study would give us knowledge of the types of problems that instructors have difficulty assessing “on the scene”. To study the influence of halo effect and fundamental attribution error in our classroom study, two groups of evaluators made weekly assessments using an instrument developed specifically to evaluate group process in detail, separately from the course grade. Since the halo effect examines an evaluator’s bias on multiple traits, the evaluators conducted assessments on

**Table 1** Group process framework: five assessment categories

Dimension	Definition
Goal setting	Making concrete plans for next steps
Progress	Fulfilling personally stated goals in a tangible way
Knowledge co-construction	Taking initiative to use knowledge or skill
Participation	Being present and active in a group work
Teamwork	Demonstrating the interpersonal skills that enable one to work in a group

five categories; goal setting, progress, knowledge co-construction, participation and teamwork (Table 1). These five assessment categories were based on previous work on evaluating PBL class work (anonymized). In “[Results from the classroom study](#)” section, we present how the hypotheses that are first introduced in “[Research motivation and hypotheses](#)” section can be explained using the data collected with an assessment model and instruments developed for this study. Based on the findings presented in “[Results from the classroom study](#)” section. “[Discussion](#)” section discusses recommendations for designing monitoring educational technology that could help mitigate the problems observed in the classroom study. Finally, “[Conclusions](#)” section presents conclusions and future directions.

## Research motivation and hypotheses

Principles from social psychology of group work showed that cognitive biases could influence assessments under conditions of limited information (Thorndike 1920; Ross 1977). We would like to explore the extent that two types of cognitive biases, namely halo effects and fundamental attribution errors, affect instructors’ assessment practices in a PBL classroom setting. Based on existing research on the two cognitive biases (Thorndike 1920; Ross 1977), it would be reasonable to expect to see these effects in classroom settings given that both biases occurred in similar context in which a person evaluates multiple individuals in a group setting with limited information. However, we do not know how strong or potentially damaging they are. Our study allows us to quantify such effects in addition to verifying their existence in PBL classrooms.

Findings from our study can be used for exploring important design questions for designing tools that analyze and monitor traces of group activities that are predictive of group process assessments (Madan et al. 2004; Rienks et al. 2006). Researchers have traditionally used qualitative and quantitative methods to assess group processes by manually coding for group processes (Meier et al. 2007; Weinberger and Fischer 2005). More recently, others have used machine learning and data mining to automatically analyze and monitor group processes (Chen 2003; Reimann et al. 2011; Rienks et al. 2006). Work focused in this way serves as a proof of concept and a good starting point for a longer term investigation in building assessment support technology that can be used in PBL design and technology courses. However, in both the manual coding and data mining approaches, evidence that this early work addresses the real issues faced by project course instructors in their assessment work is lacking. In particular, it is unclear whether the group processes addressed by researchers correspond to the most important ones that instructors would use in a classroom environment. Therefore, our study investigates the questions of “where” instructors have trouble in terms of conducting assessments in actual PBL classroom settings by using principles from social psychology.

The first principle from social psychology that is of interest, namely the halo effect, is a cognitive bias leading people to perceive another person’s traits in a way that is consistent with previous impressions of other traits (Thorndike 1920). The halo effect was first documented by Thorndike when he observed that supervisors seemed unable to rate their subordinates independently on different aspects of character. Other situations where it has been documented are when commanding officers are rating their soldiers, when a boss is evaluating employees (Beehr et al. 2001), when a customer is evaluating sales people (Lambart et al. 1997), or when a student is evaluating an instructor (Becker and Cardy

1986; Feeley 2002). However, documentation of the halo effect in situations where an instructor is evaluating students has not been explored yet.

However, as in previously documented research on the halo effect, instructors are in a situation where they have to evaluate students based on their limited interactions with them. In studies of PBL assessments, instructors have demonstrated the need for assessing students and their group work in various dimensions of assessment categories such as those shown in Table 1. Yet, instructors may be strongly biased to rate students similarly on different assessment categories (Gopinath 1999), including both those directly observed and those predicted, even when their desire is to differentiate between these categories. To study the existence and the influence of the halo effect in a classroom environment, we compare instructors' assessment in various assessment categories to the assessment of observers, who have direct experience observing group work, and propose the following hypothesis:

**Hypothesis 1** Group work instructors who are not direct observers of group work make more errors of judgment consistent with what is predicted by the halo effect.

In verifying that our data supports this hypothesis, we gain deeper insights into how the halo effect plays out in assessment of group work and explore whether having more of a direct view into group work can lessen the halo effect. In addition, since we evaluated group work on five types of assessment categories in our classroom study, we can observe which types of assessment categories are dominant, in the sense of biasing the perception of other types of categories. For example, if an instructor has a strong perception of the knowledge co-construction category, our hypothesis predicts that other categories, such as group progress, would be affected by that perception. With these types of insights, it is possible to predict exactly which types of errors of judgment will be made or in what circumstances in our assessment context.

The second principle that is associated with judgments made from limited information is the fundamental attribution error. The fundamental attribution error is the frequent tendency for people to underestimate the effect of context and to overestimate personality-based influences when explaining behaviors observed in others (Ross 1977). Although limitations and difficulty in determining the accuracy of the fundamental attribution error have been found (Harvey et al. 1981), it has been documented across multiple environments and is generally considered to be a robust phenomenon. For example, Cook and Klumper (1999) documented the effect of the fundamental attribution error regarding perception of leadership. They argue that although an individual can command authority and demonstrate leadership skills in certain circumstances, the skills do not necessarily transfer to other situations. However, people tend to underestimate the effect of context in which authority is exhibited and overestimates that effect as the individual's inherent leadership quality. Similarly, others have shown the effect of the fundamental attribution error where the Board of Probation and Parole evaluated parole cases for offenders (Carroll 1978) and students evaluated teachers (Kelsey et al. 2004; McPherson and Young 2004). However, as with the halo effect, the fundamental attribution error has not been documented in the context of an instructor evaluating students and student groups.

As in previously documented research, we expect instructors to underestimate the impact that their presence has on the behavior of their students in the context of their direct experience with them, and thus to make strong assumptions about how they behave in the group work setting that don't carry over into that context. More specifically, instructors are vulnerable to the fundamental attribution error because they have a limited vantage point where they only see the students in the context of class sessions or check-point meetings

where students put on their best behavior. An observer who attends group work sessions would be less subject to the fundamental attribution error with respect to judgments about that group work because he sees the students in the actual work context. Because the fundamental attribution error predicts that “the further removed someone is from directly observing behavior, the greater their errors of judgment should be”, we would expect to see higher correlation between the assessment of an individual and the group that he belongs to when comparing the correlation of the instructor’s assessments to the observer’s assessments. Thus, we arrive at a second hypothesis:

**Hypothesis 2** Group work instructors who are not direct observers of group work make errors of judgment consistent with what is predicted by the fundamental attribution error.

Examining our hypothesis on the fundamental attribution error would provide understanding on whether this cognitive bias occurs in PBL assessments and if so, to what degree. In particular, if we observe a high correlation between a student’s assessment score and the score of his group, developers of assessment tools can help instructors lessen the bias by providing methods for comparing the two assessment scores. Beyond seeking evidence to support or refute these hypotheses, the concrete and multi-dimensional operationalization of perspective on group work we develop and employ in this work allows us to measure the extent of errors in judgment relative to different types of assessment goals. Thus, this investigation has the potential to yield rich insights into the inner workings of group work instructors’ assessment practices.

## Method for the classroom study

We tested our hypotheses through a correlational analysis of data collected in a project course. This course was a graduate level engineering project course, in which 22 students worked on one collaborative project sponsored by an industrial client. Because the goal of the course was to have students experience a project cycle from start to finish, the instructor’s role was to guide the students in terms of the process of a project, rather than to focus on the end product. Therefore, the course evaluation was formative in nature. Four subgroups of students were formed in order to carry out the project. The semester long course was divided into three phases, each lasting 5 weeks. Each phase consisted of the general process required to complete a project, starting from planning (phase 1) to implementation (phases 2 and 3). Subgroups changed for each phase, allowing students to work with most of the students in the class.

For the assessment data, two types of data were collected: namely, observer assessment scores and instructor assessment scores. Observers were two members of our research team who observed weekly student group meetings related to the course project. The observers had experience as either an instructor or a teaching assistant in graduate level project courses. The observers introduced themselves as “flies-on-the-wall” and took measures to avoid influencing the natural group work phenomenon in such a way that might invalidate our observation. To collect data for this analysis, we asked the observers and instructors to make weekly evaluations of students in the five areas of the assessment framework (Table 1). These five assessment categories, which were established through a previous interview study (anonymized), are qualities of group work that the instructors of multiple PBL courses identified as being important. Although the course consisted of three phases, the most usable data was collected during phase 2. Data collection during phase 1 was used

for calibrating observer and instructor instruments. Data collection for observers during phase 3 didn't occur because the students didn't hold formal structured group meetings at the end of the semester. Therefore, the instructor and observer assessment data used in this paper is based on data collected during phase 2.

### Instrument for measuring observer perspective

The first instrument we developed was for use with direct observers of project group meetings. For each of the five group process dimensions displayed in Table 1, we studied the transcript data from a previous interview study with PBL instructors (anonymized). From that data we identified different questions that instructors used in association with each of the five dimensions. Using those questions as a foundation, we then constructed two to three statements that described the positive and negative student behaviors associated with each dimension at group and individual levels. For example, a statement for the knowledge co-construction dimension at the group level was "Did the group share information that is meaningful which will advance the project, or is the information superficial", and at the individual level was "Did the student present new ideas or solutions for problems being discussed during the meeting?" Table 2 shows example statements for all five dimensions as a reference.

The observers made assessments by responding to these statements associated with the five dimensions with yes or no answers. For example, for the knowledge co-construction category at the individual level, one of the statements is "Did the student present new ideas or solutions for problems being discussed during the meeting?" If the student presents new ideas or solutions, the observer would assign a "yes". However, a "no" would be assigned if the student only pushes old ideas without incorporating alternative options. Similarly, for the participation category at the individual level, one of the questions is "Does the student seem engaged in the meeting by giving full attention to the meeting itself?" The observer would assign a "yes" to the question if the student is being attentive to the meeting. On the other hand, if the student is not giving full attention to the meeting, such as checking his mobile device too frequently or not participating in the discussion, a "no" would be assigned to the question. Based on the pattern of these yes and no answers of the statements

**Table 2** Sample questions used for the five assessment categories

	Group level	Individual level
Goal_setting	Did the group discuss things to do/set goals for the next meeting?	Is the student suggesting next steps (plans, high level steps) for himself or for the team?
Progress	Was there a discussion on what the group had accomplished since the last meeting?	Where there items that the student finished during the past week?
Knowledge co-construction	Did the group share information that is meaningful which will advance the project, or is the information superficial?	Did the student present new ideas or solutions for problems being discussed during the meeting?
Participation	If work is done during the meeting, i.e. making a presentation, are all of the students doing work?	Does the student seem engaged in the meeting by giving full attention to the meeting itself?
Teamwork	Is everyone's opinion taken seriously without being ignored?	Did the student respect others' opinions by allowing them to speak/respond?

for each category, observers then assigned an overall score with a number between  $-2$  and  $2$ . The range has both negative and positive numbers so that the scorers can easily map negative behaviors to negative scores and positive behaviors to positive scores. The five point range is also convenient in terms of its correspondence with typical grading scales (e.g., A–F). For the observer instrument, two coders validated the five assessment categories and achieved a kappa of  $0.90$ .

The observers evaluated each student on the five assessment dimensions each week. In addition, an average of these five scores was computed for comparison with the course grade, which is a single score encompassing all five dimensions. Because students did not hold formal meetings during phase 3, observer data during this phase could not be collected. Therefore, we only have phases 1 and 2 scores for observer assessment scores.

### **Instrument for measuring instructor perspective**

Once the instrument for quantifying the observations of direct observers based on the five assessment categories was developed, we constructed an isomorphic instrument to be used by instructors to track their assessments of the students throughout the semester. The questions were completely isomorphic to those on the observer instrument. For example, for the sample statement in knowledge co-construction given in the previous section, the isomorphic question in the instructor instrument was “Did the student present new ideas or solutions for problems?”. Note that the only difference with this sentence is the removal of the phrase “being discussed during the meeting” at the end of the isomorphic question in the observer instrument (Table 2).

Each week, the instructor evaluated each student on the five assessment dimensions, separately from the course grade. These instructor assessments are based on classroom observations, which occur twice a week, as well as weekly meetings, which the instructor has with each student group. Because the instructor evaluation started in phase 2, we have phase 2 scores and phase 3 scores for each of the five dimensions on a weekly basis. These weekly scores were then averaged for each phase. The instructor assessments were based on instructors’ observations as well as weekly instructor-group meetings, rather than actual student group meeting observations. Again, the averages for each phase were also computed for instructor scores.

## **Results from the classroom study**

The instruments described in the previous section allowed us to compute through a correlational analysis the degree of match or mismatch between instructor, observer, and student perspectives on relevant judgments about the quality of student participation in their project groups.

### **Hypothesis 1: Group work instructors who are not direct observers of group work make errors of judgment consistent with what is predicted by the halo effect**

We can evaluate the extent to which hypothesis 1 is consistent with our data by comparing correlations across dimensions of the observer scores and the instructor scores. Previous studies that demonstrate the halo effect have reported different conceptual and operational



definitions for measurement of this cognitive bias (Cooper 1981; Becker and Cardy 1986; Balzer and Sulsky 1992). This confusion is partly due to the two different types of measurements that Thorndike used in his original study (Thorndike 1920). Therefore, we measured and reported both types of correlations that Thorndike used in his original study: (1) intercorrelations among specific performance dimensions and (2) correlations between overall ratings with specific performance dimensions. The intercorrelation score is computed using correlations among the differing dimensions, in our case the five assessment categories shown in Table 1. Because the halo effect is the tendency to rate different performance dimensions similarly, the intercorrelation score from a person that is more affected by this cognitive error would be higher. In our classroom study, because direct observers do not themselves have an instructor-student relationship with the students in the groups they observe, they may be less likely to strongly associate the separate characteristics together. Second, direct observers have first-hand experience observing the students in their groups. Taken together, we hypothesized that the direct observers may be less susceptible to the halo effect, which means that they should be able to differentiate across the various assessment categories better than instructors do. If so, the correlation between those categories of observer scores should be lower than those computed from instructor scores.

Table 3 shows descriptive statistics for the five assessment categories, and Tables 4 and 5 shows correlations scores between the categories for the observers and the instructor scores. Overall, the correlations in of the instructor's scores (Table 5) are higher than those of the observers' scores (Table 4) as suggested by hypothesis 1. More specifically, for the observers' scores, the two significant correlations are between the categories of goal setting and participation, and knowledge co-construction and participation. Comparing the correlations for goal setting and participation between Tables 4 and 5, one can see that the score for the observers is lower at 0.589 ( $p < 0.05$ ) compared to the score for the instructor, which is 0.756 ( $p < 0.01$ ). Similarly, for the categories of knowledge co-construction and participation, the correlation for the observers is lower at 0.717 ( $p < 0.01$ ) compared to that of the instructor at 0.853 ( $p < 0.01$ ). Thus, the data supports that the instructor has more difficulty in differentiating among the various assessment categories, which may be explained by the presence of the halo effect.

The second type of measurement that is indicative of the halo effect is the correlation between overall ratings with specific assessment categories. Thorndike reported that the overall impression of a person and the specific performance dimensions were in the range of 0.51–0.64 (Thorndike 1920). In this project course, the instructor assigned mid-course grades for each of the three phases, separately from the weekly measurement. The mid-course grades are based only on the project group work with an emphasis on the process rather than the product for the course. Peer assessment is not considered for the mid-course

**Table 3** Descriptive statistics for the five assessment categories

	Observer (N = 14)			Instructor (N = 22)		
	M	CI ± 95 %	SD	M	CI ± 95 %	SD
Goal_setting	3.34	2.81–3.87	0.92	4.32	4.04–4.60	0.62
Progress	3.18	2.96–3.39	0.37	4.33	4.06–4.59	0.59
Knowledge	3.63	3.07–4.18	0.96	4.15	3.93–4.37	0.49
Participation	4.63	4.21–5.04	0.71	4.37	4.06–4.68	0.70
Teamwork	5.00	–	0.00	4.49	4.29–4.69	0.44

**Table 4** Correlations for the five assessment categories of phase 2 observer scores

	Goal_setting	Progress	Knowledge	Participation	Teamwork
Goal_setting	1				
Progress	0.118	1			
Knowledge	0.273	-0.067	1		
Participation	0.589*	0.127	0.717**	1	
Teamwork	_ <sup>a</sup>	_ <sup>a</sup>	_ <sup>a</sup>	_ <sup>a</sup>	1

<sup>a</sup> Correlation could not be computed due to insufficient variation of teamwork data values

\*  $p < 0.05$ , \*\*  $p < 0.01$

**Table 5** Correlations for the five assessment categories of phase 2 instructor scores

	Goal_setting	Progress	Knowledge	Participation	Teamwork
Goal_setting	1				
Progress	0.857**	1			
Knowledge	0.765**	0.714**	1		
Participation	0.756**	0.718**	0.853**	1	
Teamwork	0.634**	0.534*	0.748**	0.823**	1

\*  $p < 0.05$ , \*\*  $p < 0.01$

grades. To compare our data with Thorndike's data, we used these mid-course grades as the overall rating and computed the correlational value between the mid-course grades along with the average of the five assessment dimensions. The correlation scores between overall ratings and average of the five dimensions for phases 2 and 3 are  $r(18) = 0.565$ ,  $p < 0.01$  and  $r(18) = 0.48$ ,  $p < 0.05$  respectively. Thus, this second measurement is comparable to the correlations mentioned in Thorndike's work, which also supports hypothesis 1.

### **Hypothesis 2: Group work instructors who are not direct observers of group work make errors of judgment consistent with what is predicted by the fundamental attribution error**

The fundamental attribution error predicts that group work instructors may make errors in judgment by attributing a student's behavior or performance to the group environment rather than to the student's own individual capabilities or motivations. Thus, a high correlation between the group's score and the individual's score could suggest that the instructor is influenced by the group's performance when assessing the individual student. To test the second hypothesis, we compared the correlation between the group score and the average of individual student scores in that group. As predicted by the hypothesis, the correlation associated with the instructor's score was higher in general compared to the one based on direct observer's score. For easy comparison, we computed the average of the

correlations across all four groups and observed that the average for the instructor is 0.40, whereas the average for the observer is 0.23.

For a more complete picture, Table 6 shows the group scores, the average of the individual scores, and the correlations between those scores for each of the four groups. Among the four groups, the comparing the correlation scores for that of the observer and instructor for group 2 and group 4 clearly shows that the observer's score is lower at 0.14 and 0.18 respectively, compared to the instructor's scores at 0.58 and 0.59. Interestingly, group 1's correlations show an opposite trend than expected. Group 1 is different from the other groups that it was composed of only three members, whereas the three other groups were composed of either six or seven members. Unfortunately, group 3's observer scores are missing, so we cannot compare the correlations. However, note that the instructor's group 3 correlation score at 0.37 is closer to that of group 2 (0.58) and group 4 (0.59) than group 1 (0.05).

Another statistic that could signal the fundamental attribution error during assessment is the variance of the individual's scores. If the instructor attributes much of the individual's performance to the group performance, then the scores of the individuals in the same group would be similar, resulting in low variance. Indeed, the variance among the students in a given group was lower for the instructor's scores ( $SD = 0.32$ ) than for the observer's assessment scores ( $SD = 0.43$ ). Thus, this second analysis also tells a similar story, which is consistent with hypothesis 2.

## Discussion

In this section, we discuss how our classroom study findings can be used for designing effective group work assessment tools. By offering instructors a deeper insight into group work that occurs outside of their direct view, assessment tools can help instructors identify situations where their concerted attention is most needed. However, prior work on

**Table 6** Group versus individual scores and their correlations for the four project subgroups

	Group				Individual			
	Grp1	Grp2	Grp3	Grp4	Grp1	Grp2	Grp3	Grp4
Observer								
Goal_setting	3.00	3.33	–	3.37				5.00
Progress	3.00	3.25	–	3.55	5.00	4.75	–	3.25
Knowledge	4.50	3.42	–	3.20	4.00	3.75	–	4.75
Participation	5.00	4.38	–	3.35	4.00	4.00	–	4.75
Teamwork	5.00	5.00	–	4.70	5.00	4.50	–	5.00
Corr group and ind					0.36	0.14	–	0.18
Instructor								
Goal_setting	4.83	4.25	4.03	3.88	5.00	5.00	5.00	4.00
Progress	4.50	4.25	4.14	4.47	4.50	4.50	5.00	4.50
Knowledge	4.17	4.33	3.93	4.53	5.00	5.00	5.00	4.50
Participation	4.67	4.58	4.09	4.23	4.50	4.50	4.00	4.00
Teamwork	4.83	4.71	4.17	4.33	5.00	4.50	4.50	5.00
Corr group and ind					0.05	0.58	0.37	0.59

automatic assessment support (DiMicco et al. 2006; Kay et al. 2006; Pianesi et al. 2008) has been largely oriented towards addressing the technical challenges rather than addressing design challenges that consider where instructors are having trouble. For example, Pianesi et al. (2008) developed a tool which automatically generates reports about participant’s behavior using audio and video annotation of meetings. Although technical challenges of generating the report itself are addressed, research on instructor needs and corresponding design choices are less explored in their work. However, since project course instructors work with project groups over an extended period of time and have regular contact with the groups, they can be susceptible to cognitive biases. Therefore, in addition to the technical challenges, design challenges that address where discrepancies of instructor view are likely to occur would benefit the assessment practices.

Deriving design principles based on the result of empirical study has previously been identified as an important research direction at computer-supported collaborative learning symposiums (Dimitracopoulou et al. 2004). Our classroom study presents data that pinpoints where discrepancies lie as well as the magnitude of those discrepancies. These findings motivate the principle that support should draw attention to differences between distinct assessment dimensions. As an illustration of one way of implementing this design recommendation, we present a design prototype referred to as the Group Assessment Platform (GRASP) in Figs. 1 and 2. GRASP is an online tool that provides data on group interaction by automatically detecting and providing the status of student groups. Such a tool would enable instructors to be more effective in identifying when students need more of their concerted attention by looking the status of the five assessment categories at both group and individual level. In other words, an instructor can “grasp” the status of multiple groups and provide guidance to groups in need.

The first hypothesis from our classroom study has shown that instructors who did not have direct access to group work were more prone to the halo effect, resulting in similar scores across differing assessment categories. Therefore, to dampen the halo effect, the assessment tool should make an explicit distinction between different assessment categories so that the instructor would be better prepared to differentiate among these various dimensions. In GRASP, we visualize each group category assessment—goal setting, progress, knowledge co-construction, participation, and teamwork—as a separate line with a distinct style, as shown in Fig. 1. Each line shows the progress along the number of weeks during a semester, which is shown in the x-axis. Use of color that highlights a negative progress in each line helps the instructor to easily detect problems and to check whether there are indeed differences between various assessment categories that he missed.

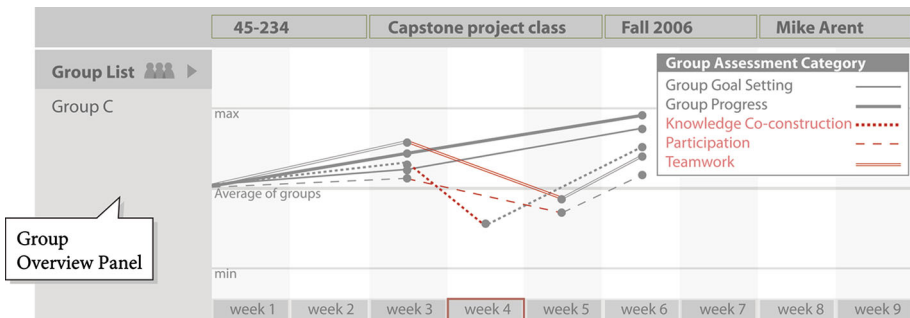
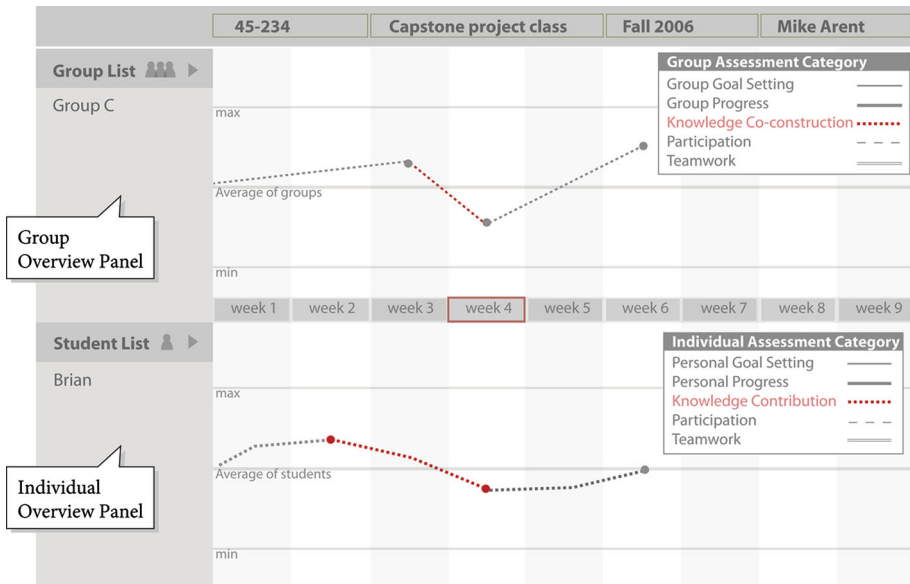


Fig. 1 GRASP interface



**Fig. 2** GRASP interface

Another important feature of GRASP that addresses our study findings, regarding the observation that group work instructors are susceptible to the fundamental attribution error, is displaying separate panels for groups and individuals. As is seen in Fig. 2, there are two different panels: the group overview panel (upper half) and the individual overview panel (lower half). Displaying both views simultaneously helps instructors compare the degree to which a group's performance might affect the assessment of an individual. For example, Fig. 2 shows that both Group C and student Brian's knowledge co-construction assessment category show a decrease in value from week 3 to week 4. Given this data, the instructor can meet Brian separately from the group to examine whether the student needs additional guidance separate from the group.

By presenting a design prototype, GRASP, we illustrated how the findings from the classroom study could be used for effective group assessments as an example. However, although GRASP has many potentials to be an assessment tool to evaluate the group work over time, its current stage as a design prototype has several limitations given the limitations of our study. First, our investigation was based on a single project class assessed by an instructor. Further study is needed to examine if our findings replicate in other types of project courses or courses assessed by more than one instructor and test the extent to which our results generalize beyond the context we have reported in this paper. Second, since phases 1 and 3 data could not be used for analysis due to administrative issues, our analysis was based on phase 2 data only. It would be interesting to observe whether the same phenomena on the two biases are observed in the beginning and the end of the course as well. Lastly, the evaluation of GRASP is needed to see whether cognitive biases affect the assessment practices and the findings could be used to improve its current design and to inform the design of other automatic assessment tools.

## Conclusions

This paper addresses important questions related to the extent and nature of mismatches in perspectives among interested parties who vary in their degree of exposure to the context in which group work occurs. The formal study we report was conducted in a project-based learning classroom. The paper describes insights gained from this investigation as well as how two principles from social psychology can be used as a lens for exploring important design questions in PBL.

We also presented how the findings from the classroom study can be applied in research that supports group assessment practices of project course instructors by providing guidelines that would enable instructors to be less prone to the psychological biases such as halo effect and fundamental attribution error. Previous research has shown that although eliminating the effect of cognitive biases is difficult, there are certain techniques to reduce the effect (Tetlock 1985; Burger 1991). Likewise, our guidelines should be tested in a formal study to see whether they can lessen the effect of cognitive biases because they were suggestions based on interpretation of our data.

In addition to the study findings themselves, we also see the potential use of the assessment instruments developed for instructor and observer assessments. The instrument used by observers is targeted for assessors who have direct access to student group meetings, whereas the instrument used by instructors can be used more broadly by group work facilitators in general in various settings. Therefore, in addition to group work facilitators in educational settings, as in the classroom study, managers in industry could potentially also use the same instrument because the five assessment dimensions are broadly defined.

Although our investigation is limited to two types of cognitive biases, researchers could adapt our methodology to test other types of psychology principles that would help in designing assessment tools. In addition, we ultimately expect to see similar needs in both an educational context and an industrial one. While the choice to conduct the work in an educational setting was made partly for practical reasons, it has advantages from the standpoint of our research focus on assessment, especially at an early stage. Conducting research that focuses on and involves evaluation could be viewed as a threat, and therefore could possibly interfere with obtaining buy-in from participants. Therefore, this work could serve as a basis for conducting future work in an industrial setting.

## References

- Balzer, W. K., & Sulsky, L. M. (1992). Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology, 77*(6), 975–985.
- Becker, B. E., & Cardy, R. L. (1986). Influence of halo error on appraisal effectiveness: A conceptual and empirical reconsideration. *Journal of Applied Psychology, 71*(4), 662–671.
- Beehr, T. A., Ivanitskaya, L., Hansen, C. P., Erofeev, D., & Gudunowski, D. M. (2001). Evaluation of 360 degree feedback ratings: Relationships with each other and with performance and selection predictors. *Journal of Organizational Behavior, 22*(7), 775–788.
- Bober, M., Sullivan, H., Lowther, D., & Harrison, P. (1998). Instructional practices of teachers enrolled in educational technology and general educational programs. *Educational Technology Research and Development, 46*(3), 81–97.
- Burger, J. (1991). Changes in attributions over time: The ephemeral fundamental attribution error. *Social Cognition, 9*(2), 182–193.

- Carroll, J. S. (1978). Causal attributions in expert parole decisions. *Journal of Personality and Social Psychology*, 36(12), 1501–1511.
- Chen, M. (2003). Visualizing the pulse of a classroom. In *Proceedings of international multimedia conference*, pp. 555–561, Berkeley, CA: ACM Press.
- Cook, M., & Klumper, D. (1999). Metacognitive, social and interpersonal skills and aptitudes in officer performance with distributed teams. In *Paper presented at RTO HFM workshop on Officer Selection, in Monterey, USA*.
- Cooper, W. (1981). Ubiquitous halo. *Psychological Bulletin*, 90(2), 218–244.
- DiMicco, J., Hoolenbach, K., & Bender, W. (2006). Using visualizations to review a group's interaction dynamics. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp 706–711, New York: ACM Press.
- Dimitracopoulou, A., Hoppe, U., & Dillenbourg, P. (2004). Interaction analysis supporting participants during technology based collaborative activities. In *Paper presented at CSCL symposium, October 7–9, in Kaleidoscope Noe, Lausanne*.
- Dutson, A. J., Todd, R. H., Magleby, S. P., & Sorensen, C. D. (1997). A review of literature on teaching design through project-oriented capstone courses. *Journal of Engineering Education*, 76(1), 17–28.
- Feeley, T. (2002). Evidence of halo effects in student evaluations of communication instruction. *Communication Education*, 51(3), 225–236.
- Gipps, C. (2005). What is the role for ICT-based assessment in universities? *Studies in Higher Education*, 30(2), 171–180.
- Gómez Puente, S. M., van Eijck, M., & Jochems, W. (2013). Empirical validation of characteristics of design-based learning in higher education. *International Journal of Engineering Education*, 29(2), 491–503.
- Gopinath, C. (1999). Alternatives to instructor assessment of class participation. *Journal of Education for Business*, 75(1), 10–14.
- Harvey, J., Town, J. P., & Yarkim, K. (1981). How fundamental is the fundamental attribution error? *Journal of Personality and Social Psychology*, 40(2), 346–349.
- Jochems, W., & Kreijns, K. (2006). Measuring social aspects of distributed learning groups. *European Educational Research Journal*, 5(2), 110–121.
- Johari, A., & Bradshaw, A. (2008). Project-based learning in an internship program: A qualitative study of related roles and their motivational attributes. *Education Technology Research and Development*, 56, 329–359.
- Kay, J., Maisonneuve, N., Yacef, K., & Reimann, P. (2006). Wattle tree: What'll it tell us? *University of Sydney Technical Report*.
- Kelsey, D. M., Kearney, P., Plax, T. G., Allen, T. H., & Ritter, K. J. (2004). College students' attributions of teacher misbehaviors. *Communication Education*, 53(1), 40–55.
- Kimbell, R. (2007). E-assessment in project e-scape. *Design and Technology Education: An International Journal*, 12(2), 66–76.
- Lambart, E., Sharma, A., & Levy, M. (1997). What information can relationship marketers obtain from customer evaluations of salespeople? *Industrial Marketing Management*, 26(2), 177–187.
- Madan, A., Caneel, R., and Pentland, A. (2004). GroupMedia: Distributed multimodal interfaces. In *Proceedings of sixth international conference on multimodal interfaces ICM104*.
- McPherson, M., & Young, S. L. (2004). What students think when teachers get upset: Fundamental attribution error and student generated reasons for teacher anger. *Communication Quarterly*, 52(4), 357–369.
- Meier, A., Spada, H., & Rummel, N. (2007). Evaluating collaboration: A rating scheme for assessing the quality of collaborative process. *International Journal of Computer-Supported Collaborative Learning*, 2, 63–86.
- Nicol, D., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218.
- Phielix, C., Prins, F., & Kirschner, P. (2010). Awareness of group performance in a CSCL-environment: Effects of peer feedback and reflection. *Computers and Human Behavior*, 26(2), 151–161.
- Pianesi, F., Zancarnaro, M., Not, E., Leonardi, C., Falcon, V., & Lepri, B. (2008). Multimodal support to group dynamics. *Personal and Ubiquitous Computing*, 12(3), 181–195.
- Price, M., Handley, K., & Millar, J. (2011). Feedback: Focusing attention on engagement. *Studies in Higher Education*, 36(8), 879–896.
- Reimann, P., Yacef, K., & Kay, J. (2011). Analyzing collaborative interactions with data mining methods for the benefit of learning. In S. Puntambekar, G. Erkens, & C. Hmelo-Silver (Eds.), *Analyzing interactions in CSCL: Methods, approaches and issue* (pp. 161–186). New York: Springer.

- Ren, Y., Kiesler, S., & Fussell, S. R. (2008). Multiple group coordination in complex and dynamic task environments: Interruptions, coping mechanisms, and technology recommendations. *Journal of Management Information Systems*, 25(1), 107–133.
- Rienks, R.J., Zhang, D., Gatica-Perez, D., & Post, W. (2006) Detection and application of influence rankings in small group meetings. In *Proceedings of eighth international conference on multimodal interfaces ICMII'06*.
- Rohde, M., Klamma, R., Jarke, M., & Wulf, V. (2007). Reality is our laboratory: Communities of practice in applied computer science. *Behavior and Information Technology*, 26(1), 81–94.
- Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 173–220). New York: Academic Press.
- Tetlock, P. (1985). Accountability: A social check on the fundamental attribution error. *Social Psychology Quarterly*, 48(3), 227–236.
- Thorndike, E. L. (1920). A constant error on psychological rating. *Journal of Applied Psychology*, 4(1), 25–29.
- Weinberger, A., & Fischer, F. (2005). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers and Education*, 46(1), 71–95.
- Wong, J., Oh, L. M., Ou, J., Rosé, C. P., Yang, J., & Fussell, S.R. (2007). Sharing a single expert among multiple partners. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 261–270. New York: ACM Press.