Respondent disengagement from a peer assessment instrument measuring Collaboration Viability

Benjamin Heslop, Elizabeth Stojanovski, Simon Iverson, Jonathan Paul & Kylie Bailey


To link to this article: https://doi.org/10.1080/22054952.2018.1437676

Published online: 12 Feb 2018.
Respondent disengagement from a peer assessment instrument measuring Collaboration Viability

Benjamin Heslop, Elizabeth Stojanovski, Simon Iverson, Jonathan Paul and Kylie Bailey

Faculty of Health and Medicine, School of Medicine and Public Health, University of Newcastle, Callaghan Campus, Callaghan, Australia; Faculty of Science, School of Mathematical and Physical Sciences, University of Newcastle, Callaghan Campus, Callaghan, Australia; Faculty of Engineering and Built Environment, School of Engineering, University of Newcastle, Callaghan Campus, Callaghan, Australia

ABSTRACT
Undergraduate group projects are intended to improve the ability of students to collaborate, a skill considered critical by many employers. If such groups become unviable, for instance due to interpersonal conflict or imbalanced effort, intervention by the lecturer may be required to alleviate this problem. The article proposes Collaboration Viability (CoVi) as a construct to indicate the degree to which a team is failing. Peer assessment instruments are designed to establish the relative contributions of team members, but are rarely used to measure emergent group-level properties, such as CoVi. Averaged per-group peer assessment ratings from 458 undergraduate engineers (forming 72 teams) demonstrated an overall weak, but statistically significant, correlation between CoVi and their group mark ($r = 0.25$, $p = 0.04$), adjusting for academic ability. A Self and Peer Assessment Resource Kit (SPARKPLUS) online instrument was used that comprised 13 questions, categorised according to whether they related to contribution, competence or teamwork. Individual questions were unevenly correlated with performance, which based upon general principles of survey design, was postulated to be attributable to poor design of some of the questions. It was further postulated that the cumulative effect of poor question design was respondent's disengagement from the instrument. Consequently, to improve the peer assessment methodology employed for this research, a novel, three-step process was recommended. First, a short, non-divisive instrument measuring peer's perception of collaboration; second, intervention by the lecturer of low-CoVi groups; and third, in recidivist groups, as SPARKPLUS currently does, application of a validated peer assessment instrument to algorithmically adjust marks.

ARTICLE HISTORY
Received 27 June 2017
Accepted 31 January 2018

KEYWORDS
SPARKPLUS; peer assessment; cognitive dissonance; disengagement; collaboration; PILAR

Introduction
A critical skill of modern employment is collaboration, which educational institutions attempt to inculcate through group projects (Cooper-Thomas and Anderson 2006; Smith, Ferns, and Russell 2014). Teams of students working on group assessments may experience difficulties from colleagues who are free-riders (non- or minimal contributors), or who possess poor technical and/or teamwork skills, any of which may cause interpersonal conflict (Levi and Cadiz 1998). To reduce dissatisfaction from students who perceive group work as unfair and/or onerous, there is a growing trend to adjust each student's mark based on reviews by their peers (Ohland et al. 2012).

Online peer assessment tools are time-effective for lecturers who have, in the past, manually compiled peer assessment ratings that had been entered on paper (Loughry, Ohland, and Woehr 2014). Triantafyllou and Timcenko (2014) conducted a literature review of peer assessment of engineering student projects, and identified SPARKPLUS, CATME and WebPA as widely accepted online tools. Whereas SPARKPLUS and WebPA expect the assessor to design their own peer assessment questions, CATME's questions are preset (Delaney et al. 2016).

The Self and Peer Assessment Resource Kit (SPARKPLUS) instrument is commonly employed in Australian university engineering departments (e.g. Daly 2014), and was the package used for this research. As with all online peer assessment tools, SPARKPLUS is intended to identify free-riders through peer assessment, and downgrade their marks accordingly; or otherwise to upgrade the marks of those who contribute strongly (Willey and Gardner 2008). Its secondary purpose is educating students based upon others' perception of their teamwork behaviours, should they be markedly lacking.
Background

When collaborating, each group member naturally makes not only assessments of each of their peers, but also of the team. If members perceive their team as fair, productive and friendly, engagement will follow (Costa, Passos, and Bakker 2014). This article terms collaboration viability (CoVi) as the basket of within-team factors that influence engagement (Costa, Passos, and Barata 2015). CoVi is similar to team viability, which measures members’ perceptions of team prospects, and achievement of affective needs, such as for belonging and liking (Hackman 1986). Unfortunately, numerous biases are exhibited by respondents when asked for their own-perception of CoVi, which reduces the effectiveness of this approach. Please note that while self-perception is accepted nomenclature, it is defined by psychology and neuroscience as examination of the self (e.g. Pfeifer et al. 2009).

Potential biases when surveying Collaboration Viability

Numerous theoretical constructs are relevant to the aspirations of CoVi, prominent among them being psychological safety, group harmony and cohesiveness. Psychological safety is the extent to which a team feels safe for interpersonal risk-taking, such as suggesting an idea or admitting a mistake (Edmondson 1999). Group harmony refers to intra-group relationship quality, characterised by attribution of benign motives when disagreement is encountered, and an overall balance between individual needs and group unity (Chen et al. 2016). Cohesiveness considers liking, task commitment and group pride (Lott and Lott 1965). Each construct has accompanying instruments designed to assess a respondent’s collaborative environment from their perception of its personal impact (Newman, Donohue, and Eva 2017).

Unfortunately, own-perception of the group may experience various biases due to a respondent’s desire for self-enhancement (Mathieu, Gilson, and Ruddy 2006). If such a respondent feels they are insufficiently celebrated, or overly disadvantaged by the group, own-perceptions of CoVi may be unduly negative (Lönnqvist et al. 2008). Even for respondents aware of their deficiencies, self-reporting bias encourages survey respondents to hide potentially embarrassing aspects of themselves. Hence, poor collaborators may divert attention from themselves by emphasising faults with the group or with specific colleagues (Donaldson and Grant-Vallone 2002).

Another potential bias occurs from a respondent’s fixed mindset, which assumes that both oneself and others are unable to change, and therefore unable to improve. Observing deficiencies leads to negative evaluation, rather than suspending judgement in anticipation that deficiencies might be rectified (Chiu and Dweck 1997). Since one’s own strengths are typically given priority when evaluating another, colleagues are more likely to be deficient (Dweck and Leggett 1988). Hence, those of a fixed mindset develop an overly pessimistic view of others’ capabilities, while those of a growth mindset gain fuller appreciation, which improves relationships (Zingoni and Corey 2016). As a result, those with a fixed mindset make poor collaborators because they negatively evaluate those possessing dissimilar skills. Yet synergistic combining of diverse, complementary skills is a major benefit of collaborating, compared to a hierarchical leader valuing strengths they personally possess, rather than subordinates’ actual capabilities (Hughes and Jones 2011; Thomas et al. 2007).

Due to these three biases (self-enhancement, self-reporting, fixed mindset), there is a genuine risk of false negatives when respondents are asked to rate CoVi (Hoendberdois 2013). It is, however, not possible to identify biased respondents in own-perception instruments because ratings are inherently subjective (Kim, Baek, and Kim 2011). There is scant research taking the alternative approach of inferring each member’s perception of CoVi from assessment of their peers, which may enable detection of bias through anomalous patterns within the group’s matrix of ratings (Montgomery 1986).

The potential weakness of peer assessment is inter-rater bias due to a personal relationship, which may cause misreporting of a peer’s ratings (Lee et al. 2013). Yet, it is unlikely that any two respondents will share the same bias against a peer, or at least not uniformly expressed within the same questions. Therefore, we contend that the matrix of ratings accumulated across all respondents in a team may allow for biases to be detected, and certain ratings removed. An algorithmic approach to automatically remove biased ratings in a matrix of ratings has not, to our knowledge, been discussed in the literature (Kennedy, Kim, and Chan 2005). The following section will explore hypotheses related to extracting, from peer assessment data, respondent’s own-perception of CoVi.

Hypotheses

This research seeks to establish whether peer assessment is a potential method of measuring CoVi. It was previously noted that each team member would independently reach a view on whether their team was performing well or poorly. However, the independence of member’s own-perception of CoVi may be compromised by intra-team processes such as emotional contagion (Barsade and Knight 2015), which allows individual affect to emerge as the affective tone of the group (Collins, Jordan, Lawrence, and Troth 2016). It is considered that the group’s affective tone will bias peer assessments – higher for positive affect, and lower for negative affect – and hence reveal CoVi.
**Improvement hypothesis**

Due to member's perceptions of CoVi being cumulatively accurate, it was considered that the average of member perceptions of CoVi would correlate with team performance (Costa, Passos, and Bakker 2014). However, average member intelligence has been found to be the strongest contributor to group performance (Bates and Gupta 2017). To lessen confounding from member intelligence, *Group's Improvement over Expectations* (GloE) is defined as the difference between the final group mark awarded, and the group average of member's academic ability, based upon their final exam mark (Devine and Philips 2001; Ellis et al. 2003). Mumford (2010) attained a strong correlation \((r = 0.49, p < 0.05)\) between the average of the group's peer assessment ratings and group performance. Therefore, we hypothesised that CoVi would also be strongly correlated with GloE (H1) (Jamalludin, Razman, and Niza 2016).

**Disengagement hypotheses**

Respondents that disengage from a peer assessment instrument will give less accurate ratings (Ferguson et al. 2016; Kennedy, Kim, and Chan 2005). Making reliable between-group comparisons of CoVi requires consistency over the cohort between respondent's interpretations of each question. Measuring CoVi may therefore be more sensitive to poorly designed peer assessment questions, compared to rating peers within the team, who may be individually down-rated via questions that are not perfectly descriptive of their behaviour. Consistent interpretations also depend upon engaged respondents who sincerely attempt answers, and hence two measures of respondent disengagement are proposed.

**Group variation in ratings**

This article hypothesises that correlation of CoVi and GloE will be moderated by respondent disengagement, itself provoked by poorly designed questions (H2). Two methods of measuring respondent disengagement are proposed: *group variation in ratings* (GVR) and *set-point tactic* (SPT). Regarding the former, we consider that an engaged respondent will provide greater variance in ratings between questions for each peer. GVR quantifies engagement through averaging the standardised difference (SD) of each of the group's sets of questions (Loughry, Ohland, and Woehr 2014). For example, in a four-person group, each respondent will have three SD's, and the average of all 12 is the group's GVR.

Between the interim and final surveys, respondents were assumed to have become more familiar with one another, increasing GVR. Alternatively, if respondents were dissatisfied with the instrument's questions for the interim survey, they may be prompted to disengage when the final survey is applied, reducing GVR (O'Neil, McLarnon, and Carswell 2015). It was therefore hypothesised that respondent capacity to assess their peers from the interim to the final survey, would not be offset by disengagement from a poorly designed instrument, translating to increased GVR in the final survey, when averaged across the cohort (H3a).

**Set-point tactic**

Disengaged respondents may also attempt to minimise their cognitive effort (Hess and Statopoulos 2013) in a behaviour that Lance, Woehr, and Fisicaro (1991) term the *categorisation model*. However, this article considered that nomenclature to be unclear, and used SPT in its place. SPT was a process postulated to begin when the respondent selects a set-point rating for a peer. The respondent then selectively lowers only a few ratings below the set-point, at which other ratings remain.

We conjecture that SPT reduces cognitive load in two ways. First, if the set-point is already positioned at a high rating, only negative behaviours or incidents of the peer need be considered (Murphy and Constans 1987), for instance regularly interrupting (behaviour), or a missed deadline (incident) (Shapira and Shirom 1980). Recalling negative behaviours imposes less cognitive load than dedicated evaluation of each question, comparing numerous positive and negative behaviours and incidents of the peer (Breugst et al. 2012; Panadero, Romero, and Strijbos 2013). Secondly, SPT may also reduce cognitive load because each saliently recalled negative behaviour or incident need only be loosely matched against one of the questions, which is a lower imposition compared to properly evaluating each question. SPT may result in higher overall ratings if the respondent cannot accurately recall negative behaviours and incidents.

We estimate adoption of SPT by each respondent through the proportion of ratings that fall within \(\pm 1\) of the mode of ratings they have entered for each peer. This is ostensibly only a small likelihood given that the range of ratings is 0–100. For example, if in a 5-item instrument the ratings were 80, 81, 80, 80 and 85, then only one rating would fall outside of the range 79 to 81, giving a SPT-score of 80%. Each respondent will enter a set of ratings for each peer, and the average of this set of SPT-scores is the respondent's SPT-mean. If the respondent's SPT-mean is over 50%, we consider it is likely they are employing the set-point tactic when responding to the peer assessment instrument.

It was postulated that an instrument with poorly designed questions would motivate higher levels of SPT, due to respondent's hesitation to apply considered thought (Strijbos and Sluijsmans 2010). For those respondents who have adopted SPT, there will be a skewed distribution of ratings below their chosen set-point, increasing average ratings for groups with high SPT. If SPT were not prevalent, rating variation would have no correlation with CoVi because ratings would be normally distributed around the mode for each set. Thus,
it is hypothesised that where SPT occurs in a majority of a cohort, GVR will be negatively correlated with CoVi (H3b).

Demographic hypotheses

Being part of an out-group, minority members fear that they will receive a smaller share of the eventual reward (Knowles et al. 2010). In this situation, any such fear is irrational since group marks are not zero-sum, yet ancestrally, they were; food and status, for instance. Since perceived status alters respondent’s anticipation of receiving a share of group success (prospects), it was postulated that minority members will be more likely to engage with the peer assessment instrument (Lievens 2001). This is supported by Gatfield’s (2006) finding that minority international students express 13% higher satisfaction with a peer assessment instrument.

Thus, it was hypothesised that groups having diversity of gender (H4a) and nationality (H4b) would exhibit higher GVR (Wittenbaum, Shulman, and Braz 2010). We considered that the same rationale did not apply to older students; who possessed seniority, and hence self-esteem, that compensated for their minority status (Gatfield 2006). Hence, it was hypothesised that age diversity would have no correlation with GVR (H4c).

Method

The focus of the study was first-year undergraduate engineering, surveying and computer science students from the University of Newcastle, Australia, who were completing a design project as part of their coursework. In teams of three to six members, 458 first-year engineering students completed a peer assessment instrument (404 males, 54 females; mean age 20.1 years, s.d. 4.6 years; and 91.9% Australian national). Students also attended weekly two-hour lecture sessions on topics relevant to the stage of the project, followed by three-hour tutorial sessions, where the students worked in their teams with input from tutors. Students were split into different tutorial sessions and design projects based on their specific degree programme.

Each student’s age, nationality and gender were collected upon their enrolment in the course. The lecturer assigned groups with the intention of minimising potential for conflict or sexual harassment (Oakley et al. 2004). Given this cohort was from a first-year engineering course, groups tended to be homogenously young, male and Australian; students were allocated in bands based on their age (25% of groups comprised students older than 21 y.o.); where possible female students were assigned so that groups were either exclusively male (60% of groups) or else had at least two female members (40% of groups); international students were distributed so that there was only one in a group (40% of groups), so as to reduce the perceived burden on the group by the Australian students, and thus hopefully increase the likelihood of sympathetic reception and interaction.

First half of semester

In the first half of the semester, each group’s focus was on analysing the project brief to identify the aims, scope, constraints and performance criteria, followed by generating potential solutions and then evaluating these against the performance criteria. There was also lecture input on project management, report writing and referencing conventions. During those first six weeks, each group submitted four small (one to two pages long) tasks worth 5% of the course mark. These were designed to encourage students to develop a group code of conduct, write effective meeting minutes, plan for the drafting and editing of their interim report, and evaluate the team’s functioning. Each group submitted an interim report in Week 5 that was worth 10% of the course mark.

Second half of semester

The second half of semester focused on more detailed triple bottom line evaluation of the chosen design (economics, environmental and social), as well as presentation skills. Each group gave a presentation in Week 10 or 11 worth 10% of the course mark, and then submitted a final report in Week 13, worth 25% of the course mark. The other 50% of course marks came from the best five of seven online quizzes, worth a total of 15%, and a final multiple-choice exam worth 35%, consisting of 70 multiple-choice questions.

SPARKPLUS online assessment

Peer and self-assessment was collected after submission of both the interim and final report. The results were used to adjust individual student marks from the group mark. Groups were tasked with producing a report regarding an engineering problem that required research and analysis to produce. The report was produced in two stages, with an interim report due Week 5, and the final report due Week 13. An interim and final SPARKPLUS online survey was submitted anonymously by each student. Students were penalised 10% of their individual mark if they failed to complete peer assessment within a week of each report being submitted.

Before attempting a survey, students received written instructions on how to use the survey interface and how the group’s data might be used to adjust their mark. Students were warned via email by the lecturer that overestimating their own contribution at the expense of others may result in their evaluations being omitted from the analysis. They were told that the total pool of marks available to the group members was fixed, and so therefore when they chose to rate another member more highly, they were in effect giving permission to
the lecturer to take some of their own marks and award them to the high-performing member of the group. They were also asked to mark ‘average’ behaviour at 70 out of a possible 100, and to use very low or very high marks for exceptionally poor or favourable teamwork behaviours, respectively.

The SPARKPLUS instrument was administered online with opportunities for both peer and self-review provided for each question. Ratings were entered on a continuous scale from 0 to 100. Zero to ten equates to well below average, while 90–100 is well above average, and students manually adjusted ratings by dragging a line within a horizontal bar.

**PILAR model of collaboration**

**Design and classification of peer assessment questions**

Peer assessment was undertaken by each student via a 13-question instrument, categorised according to whether each question measured workload and effort (questions 1–3), technical quality and proficiency (4–6) or team work (7–9). The first nine questions were not able to be altered for this research as they had been previously used for the engineering course. To gain additional insight, we appended four additional questions (10–13). Their design was inspired by the PILAR model of collaboration, which is proposed to encapsulate extensive social psychology theory, underlying own-perceptions of CoVi (Heslop, Stojanovski, Paul and Bailey forthcoming). PILAR asserts that prospects, involved, liked, agency and respect perceptions of collaboration, or Pillars, have equivalent importance in member’s assessment of CoVi (Table 1). A similar equivalency is assumed by the CATME peer assessment instrument, the dimensions of which are similar to PILAR’s perceptions (Ferguson et al. 2016). Hence, for each team, CoVi was calculated as the average of all respondent’s ratings of all peers.

While the first six questions were considered to sufficiently elaborate prospects and respect, four additional questions were necessary to more fully elaborate involved (10, 12), agency (11) and liked (13) (Table 2). Question three was not considered to align with a Pillar since individual punctuality may depend upon factors other than commitment to the team. For example, an organised character, cultural background and schedule conflicts (Back, Schmukle, and Egloff 2006; Basu and Weibull 2002).

**Psychological mechanisms of respondent disengagement**

Respondent disengagement aligns with the involved Pillar, which also includes voluntarily supplying requested information. Based upon this, we propose three causes of a respondent’s reduced involved perception. Firstly, if a respondent’s team is functioning effectively, prospects are perceived to be high, reducing necessity of peer assessment. Secondly, if poorly designed questions fail to address important aspects of a peer’s behaviour, or incorrect assumptions are made, then the respondent has less respect for the survey designers. Thirdly, questions with multiple interpretations imply that a given rating may be misinterpreted by the assessor, which reduces the respondent’s perceived agency. Hence, PILAR was postulated to elucidate the psychological mechanisms of respondent disengagement, and to classify the dimensions of CoVi.

**Removal of outlier teams**

Final, rather than interim, peer assessment ratings were selected for analysis, since increased interpersonal familiarity developed over the semester (Goodman and Leyden 1991) was presumed to improve accuracy of peer assessment. Of the 93 teams in the final cohort, only 72 were included in this analysis. The 21 other teams were removed as potential outliers. Of these, six teams with three members were removed since their dynamics were considered potentially distinct from larger teams, and their matrix too small for analysis. Due to potential confounding from that member’s poor contribution, 13 teams that had members who did not participate in peer assessment were also removed (it is presumed those members had a general disinterest in the course, usually preceding their withdrawal, since otherwise a 10% penalty in that individual’s report mark was incurred for failing to complete the survey). Two teams were removed due to collaboration breakdown that was evident from pairs of respondents exchanging ratings lower than 10.

**Results and discussion**

Pearson’s correlation coefficients were used to assess the strength of bivariate linear relationships between the constructs. A statistically significant correlation was found between GIoE and CoVi scales ($r = 0.25$, $p = 0.04$). This result implies that peer assessment is a viable method of measuring CoVi (H1). However, the correlation was relatively small in magnitude, suggesting that our peer assessment instrument may lack validity.

<table>
<thead>
<tr>
<th>Pillar</th>
<th>PILAR perception (Pillar) definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prospects</td>
<td>Your opinion of whether the group will succeed, and if so, whether you will receive your anticipated share of that success</td>
</tr>
<tr>
<td>Involved</td>
<td>Your willingness to cooperate with colleagues, either providing or receiving assistance in the form of knowledge, and physical aid</td>
</tr>
<tr>
<td>Liked</td>
<td>Your sense of popularity and security based upon colleagues’ warmth and affection towards you</td>
</tr>
<tr>
<td>Agency</td>
<td>The permission you feel to suggest change to the group’s norms, processes, task allocation and strategy</td>
</tr>
<tr>
<td>Respect</td>
<td>Your opinion of a colleague’s task-relevant competence, and general trustworthiness</td>
</tr>
</tbody>
</table>

Source: Heslop, Paul, Drew, Bailey, and Stojanovski forthcoming
Pearson's correlation between the average of each team's respondent's ratings for a question, and its GloE.

Table 3. Pearson's correlation between the average of each team's respondent's ratings for a question, and its GloE.

<table>
<thead>
<tr>
<th>No.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>0.20</td>
<td>0.19</td>
<td>0.19</td>
<td>0.21</td>
<td>0.17</td>
<td>0.29</td>
<td>0.27</td>
<td>0.25</td>
<td>0.25</td>
<td>0.23</td>
<td>0.21</td>
<td>0.23</td>
<td>0.27</td>
</tr>
<tr>
<td>p</td>
<td>0.00</td>
<td>0.01*</td>
<td>0.08</td>
<td>0.16</td>
<td>0.01*</td>
<td>0.02*</td>
<td>0.03*</td>
<td>0.03*</td>
<td>0.06</td>
<td>0.07</td>
<td>0.06</td>
<td>0.02*</td>
<td></td>
</tr>
</tbody>
</table>

*p < 0.05.
There are three results from these data suggesting that GVR varied primarily by engagement. Firstly, GVR decreased 14% between interim and final time periods, while CoVi remained constant. This is consistent with the postulated reaction to student’s recalling poorly designed questions from the interim survey, and subsequently disengaging for the final survey. Secondly, 50% prevalence of SPT in the final survey suggests the presence of respondent disengagement. This represents a 39% increase in respondent disengagement compared to the interim survey, for which SPT prevalence was 36%. This 39% increase is independently confirmed by a 14% decrease in GVR, offset by a 26% increase in variation due to increased familiarity (O’Neill, McLarnon, and Carswell 2015).

The third result supporting GVR as a measure of disengagement was higher levels of variation found in teams with minority respondents. If students disengaged after the interim survey, poor question design may have been at fault. Only 5 out of 13 questions correlated significantly with GloE, with the remaining 8 very weak and non-significant. In the following section, characteristics differentiating poorly from well-designed questions are examined.

**Making survey questions meaningful**

Based upon the PILAR model, three effects of poor question design were previously proposed. Two effects were contended to generate disengagement: the respondent feeling less *respect* for the survey designer due to questions containing improper assumptions, or not being well targeted; and the respondent feeling less *agency* because questions have multiple interpretations. A third cause of reduced accuracy was *opacity*, due to lack of familiarity with the peer, rather than respondent disengagement (Fowler 1995; Montgomery 1986).

We further contend that a respondent experiencing *cognitive dissonance* (Freedman 1965) may strengthen the above disengagement-related effects, due to subjective emotion aroused by the question being contradicted by the objective answer. Much of the literature on peer assessment treats affective reactions as unwelcome variation (Cardy and Dobbins 1986; Koh, Hong, and Seah 2014), yet emotions represent valuable information. In extremis, interrater affect becomes biased, but within normal ranges, emotions provide guidance for collaborating with peers, such as predicting which peer best suits a task (Costa, Passos, and Barata 2015). For example, affective preferences are evident during selection of the successful interview candidate, or considering who to invite into a sports team (Greenberg et al. 2010). For questions that were not significantly correlated to GloE, characteristics are explored that potentially; reduced respondent’s *respect* and *agency*, increased opacity, and exacerbated cognitive dissonance (Hess and Stathopoulos 2013).

**Question one.** How much work and effort has the person put into the project? Have they done their fair share and pulled their weight?

The first problem with Question one is it contains numerous parts, that may each conceivably be answered differently (Fowler 1995). For example, a motivated peer who, due to insufficient expertise, could not contribute meaningfully to the team. Oakley et al. (2004) suggest two methods of assessing a peer’s effort: numerous questions, each related to narrowly defined behaviours; or a single question with numerous rating levels, each given a descriptive anchor (Debnath 2015).

**Question three.** Did the person regularly attend team meetings on time?

Poor attendance at meetings may be due to unavoidable schedule conflicts, which for this cohort may involve work or study commitments (Oakley et al. 2004). Yet even without attending meetings, a peer’s contribution may be substantial if they deliver effectively upon set tasks. Conversely, colleagues who routinely attend meetings may be non-contributors, or even disruptive (Channon et al. 2016). In all these situations, Question three may cause cognitive dissonance, generating respondent disengagement.

**Questions four and five.** How good is the quality of the person’s research and their understanding of what they have read? How good are the person’s analytical and problem solving skills?

Questions four and five create uncertainty regarding whether peer incompetence is caused by a lack of effort or lack of ability. If the cause is poor ability, students may feel guilty rating a peer poorly if they had contributed usefully on limited, simpler tasks within their capability. Alternatively, the peer might have misunderstood the task, which led them to a poor understanding of the problem. All of these three potential causes objectively justify a poor rating, yet only lack of effort might trigger the student’s instinctive avoidance mechanism of a poor collaborator, and so avoid cognitive dissonance.

**Question six.** How good are the person’s reporting writing and editing skills?

Question six is similarly worded to questions four and five but significantly correlated to GloE, perhaps because it relates to report writing, the quality of which had a strong, proximate effect on marks awarded. Nevertheless, question six may still have caused disengagement for reasons previously discussed.

**Question ten.** How well did the person assist others when asked?

Another potential cause of cognitive dissonance is a *loaded question* that may be operating upon an incorrect assumption, which does not apply for all colleagues. For example, Question ten makes an assumption that a peer was asked to assist, yet this may not be the case when incompetent or disliked. Nevertheless, if such a
peer were to be asked to assist, compliance may have inevitably followed, leading to an affirmative answer. Regardless, the student feels this peer is an undesirable collaborator, so a positive rating will create cognitive dissonance.

**Question eleven.** Did this person express their opinions with confidence?

Question eleven assumes that confidently expressing an opinion is a universal good, but consider a peer who arrogantly expresses their opinion, or worse, if that opinion were also incorrect?

**Question twelve.** Did the person constructively challenge other peoples’ opinions?

Since having one’s opinion challenged is rarely pleasant, the peer responsible may be affectively perceived as an undesirable collaborator, and so this question may generate cognitive dissonance. Of course, some have the capacity to objectively appreciate their colleague’s feedback, but for instance, being contradicted is always unwelcome for those with low goal orientation (VandeWalle and Cummings 1997).

**Comparison with similar research**

Peer assessment instruments typically ask respondents to rate whether, or to what extent, a peer demonstrates certain behaviours or traits (Kusano, Conger, and Wright 2015). In examining the literature, only Mumford (2010) was found to use peer assessment instruments to measure CoVi, and the article has only been cited once. A recent systematic review of team work pedagogy in higher education found that peer assessment was regularly featured, but never employed to measure team-level properties such as CoVi (Riebe, Girardi, and Whitshed 2016). By comparison, Balkundi and Harrison (2006) reviewed social network analysis (SNA) research and discovered that density of both expressive and instrumental ties (respectively, high Liked, and high Involved) were commonly investigated for their correlation with perceived team viability, and actual team performance.

We investigated whether collaboration viability (CoVi) can be meaningfully predicted from peer assessment data. In partial agreement with Hypothesis 1, CoVi calculated as per-group mean of ratings, was weakly correlated to the group’s collaborative performance, which was measured as GloE above individual academic ability. As the only comparable experiment identified in our literature review, Mumford (2010) attained a strong correlation, but with different conditions. Rather than 13 questions, Mumford (2010) asked 30 peer assessment questions based around 10 team roles, each consisting of 3 questions pertaining to role-related behaviours. Further, students were informed that the survey was for research purposes only, and their marks would not be adjusted.

Out of the 10 roles, only Mumford’s (2010) cooperator was not correlated with group performance ($r = 0.01$). In considering respondent disengagement as the cause, Mumford et al. (2008) stated that the cooperator role is less relevant to egalitarian teams. This concurs with our finding that poorly designed questions will deliver reduced correlation with GloE. However, the cumulative effect of many such questions is to lower overall respondent engagement in the instrument. Mumford’s 30 questions appear to not have caused respondent disengagement, and hence overall correlation was strong.

One further cause of disengagement is respondents’ providing a larger number of ratings for each peer within larger teams. For this, the CATME instrument was progressively shortened from an initial 87 questions to 32, and then finally to a 5-item behaviourally anchored (BARS) instrument (Ohland and Loughry 2006). As was observed with our data, lower correlated questions tended to have little effect on the overarching CoVi. Greater inaccuracy from poorly designed questions that create random inaccuracy in ratings, appear to have had a cancelling effect when averaged across other questions within each group. However, since Mumford’s data showed three distinct levels of correlation ($r = 0.0$, 0.2 and 0.4), it appears that respondents answered all three questions within each role uniformly. Hence, respondents may be effectively answering only 10 questions by duplicating their responses within the constituent three; but, as with our data, less-correlated questions are noise that averaging can remove.

Hence, it is postulated that relatively few (five) well-designed peer assessment questions will measure CoVi effectively (Heslop et al. 2017). While we have calculated CoVi as an average, further measures may be useful, such as between-peer variation calculations designed to discover free-riding (Harding 2017). Once a team with poor CoVi has been identified, intervention may be applied. Oakley et al. (2004) found that reflective listening fosters mutual appreciation of a colleague’s perspectives, and is sufficient to change behaviour in a clear majority of cases. We therefore recommend that any team found to have minimal CoVi be directed towards an in-person intervention by an independent authority figure, such as the lecturer. After such an intervention, the benefit of a short instrument is that it may be reapplied with little time commitment from respondents. In those few cases where collaboration remains unviable following intervention, the data collected (perhaps over numerous applications) may be used to adjust marks.

When adjusting marks, it may not be clear how fault may be apportioned, since respondents may be experiencing bias, or attempt to game the marking system. As previously noted, an algorithmic method of interrogating the matrix of ratings would be beneficial, to avoid the lecturer making subjective judgements. Kennedy, Kim, and Chan (2005) investigated the probabilities of any
two questions being correlated using bootstrapping and permutational methods. We believe this method can be extended by varying the weight on each team member’s data, to understand which set of weights gave the maximum correlation between questions, and therefore to the postulated CoVi. Students with lowest weights on their data are more likely to be experiencing interrater bias, or deliberately gaming the instrument. Their ratings can be factored less when marks are adjusted using existing means, such as SPARKPLUS SPA Factor (Willey and Gardner 2008).

However, a potential drawback of peer assessment is that critiquing peers may make respondents hesitant to be accurate, and potentially seed division (Montgomery 1986). Another benefit of Mumford’s approach is specifying team roles, which compared to behaviours reduce the severity of peer assessment since peer’s role competence is expected to vary. We therefore recommend establishing a method to assess CoVi that reduces emphasis on student’s native ability, instead assessing contribution relative to their capacity. Even less potentially divisive is asking respondents to evaluate their peer’s perceptions of CoVi, with questions for instance formulated via the PILAR model (Table 2). Another advantage of such indirect peer assessment of CoVi is that the instrument is more difficult to game, since a peer’s perceptions of CoVi may in fact be accurate, even if negative. However, once poor CoVi has been established and intervention has failed, a direct peer assessment tool may be required to adjust grades, formulated upon existing tools, such as CATME or Mumford’s team roles.

**Conclusion**

Identifying respondent disengagement remains a standing problem in the literature. Other research agrees with our GVR construct in assuming that respondent’s within-set ratings variation indicates respondent engagement (Kennedy, Kim, and Chan 2005). SPT is a novel approach in the context of peer assessment, and it received some validation by demonstrating identical levels of disengagement to GVR, when the effect of familiarity was included. However, deployment of SPT to measure engagement is limited to continuous scales, rather than discrete, such as Likert or BARS. In such cases, we recommend measuring the respondent’s online completion time, since rapid completion of the instrument may imply disengagement (Fan and Yan 2010).

Collated data demonstrated the potential for peer assessment to measure CoVi in student teams, but it should ideally be brief, engaging and non-polarising. Considering our finding that minority students more strongly engage in peer assessment, it seems an important research direction given the growing prevalence of student project assessments, potentially from gender-imbalanced courses, and including international students. While the use of peer assessment to measure Collaboration Viability has been anticipated by numerous authors, and in one case a research plan lodged, it remains an under-studied topic of research (Koh, Hong, and Seah 2014; Ohland et al. 2015; Salas et al. 2015).

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**Notes on contributors**

Benjamin Heslop, MPhil, BEng (Systems) is a PhD candidate in the field of social psychology at the University of Newcastle, Australia. He has earned a Master of Philosophy examining entrepreneurial collaboration, and a Bachelor of Engineering, both at The Australian National University. Benjamin is investigating collaboration using engineering systems modelling. His research combines social psychology research into a single framework, consistent with evolutionary theory. Benjamin hopes to use his research to influence government policy, and to educate students and employees.

Elizabeth Stojanovski, PhD, BSc (Maths, Hons), is a lecturer in statistics and biostatistics, and has a research focus on longitudinal relationships between life events and both physical and mental health. Elizabeth adapted and extended statistical methods including Bayesian, longitudinal, structural equation and meta-analysis models. Elizabeth consults with academic colleagues, and postgraduate students, in analysing research data and preparing manuscripts. She is a reviewer of NHMRC grants, and for the Medical Journal of Australia.

Simon Iveson, PhD, BEng (Chemical, Hons, University Medal), Grad Dip (Christian Studies), is a research associate and sessional academic at the University of Newcastle. His research focus is on coal and mineral processing. He teaches a range of undergraduate and postgraduate courses within the School of Engineering. In his spare time, he enjoys playing underwater hockey, LEGO, and board games.

Jonathan Paul, PhD, BSc (Hons) is an emerging leader within the field of reproductive medicine working toward the development of novel strategies for clinical intervention during major obstetric complications. As a scientist developing projects focused toward clinical translation, Jonathan has a practical interest in mechanisms that underpin effective collaboration, and how academic researchers and clinicians can collaborate effectively to optimise clinical translation of laboratory-based research.

Kylie Bailey, PhD, B.A. (Psych) M.Psych Clin, PhD (Psychiatry), is a senior clinical psychologist who has over 20 years clinical experience. Kylie’s expertise is with addiction, childhood trauma, self-harm and posttraumatic stress disorder, and she is also currently conducting research on combined depression, alcohol and posttraumatic stress disorder in adults. Kylie also teaches postgraduate addiction and mental health courses, and is employed as a clinical manager.

**References**


B. HESLOP ET AL.
of Collaboration Viability.” Manuscript Submitted for Publication.


