Journal of Problem Based Learning in Higher Education

VOL. 6, NO. 2, 2018 – Page 58-70
10.5278/ojs.jpblhe.v6i2.1992

Random Allocation of Students into Small Groups in Problem-Based Learning can Create Significant Between-Group Variation During the Assessment Process

Timothy Work and Yves Mauffette *

ABSTRACT

In problem-based learning large cohorts of students are divided into smaller groups that pursue learning objectives with separate instructors called tutors. This presents challenges for tutors tasked with providing similar educational experiences and assessment of multiple groups of students. Here we evaluated between-group variation in test scores that are attributable solely to the random sampling without replacement process used to form smaller groups. We then compared this with the actual between-group variation in test scores in a university-level zoology class over 4 years. We found the variation attributable exclusively to group formation accounted for a 14.4-16.2 point differential between groups whereas differences in empirical test scores attributable to group formation and other factors such as tutor capacity and group dynamics ranged from 12-18 points and rarely exceeded the variation inherent solely to group formation. This implies ad-hoc strategies for reducing variation between groups at the assessment phase will have limited success.

Key words: Problem-based learning (PBL), between-group variation, assessment, grades

INTRODUCTION

A number of disciplines have incorporated active learning approaches, either partially or completely into their curriculums (Prince 2004; Galand & Frenay 2005). Problem-based learning (PBL) is one such active learning approach that has been present for a number of years in medicine (Barrows 1996) and engineering (Mills & Treagust 2003; Prince & Felder 2006) and has become more common in the fields of science (Akçay 2009; Mauffette & Poliquin 2001) and social sciences (Heycox & Bolzan 1991). Benefits of PBL approaches include

* Timothy Work, Université du Québec à Montréal, Department of Biological Sciences, Canada
Email: work.timothy@uqam.ca
Yves Mauffette, Université du Québec à Montréal, Department of Biological Sciences, Canada
Email: mauffette.yves@uqam.ca

promotion of soft skills (Bell 2010) as well as long-term retention of course material (Strobel & van Barneveld 2009). However, major concerns remain as to how students are assessed in PBL learning environments (Azer 2003, Macdonald & Savin-Baden 2004, Macdonald 2005). Baden (2004) expressed concerns that students in the PBL context may feel that their learning is unrewarded and that working in groups is undervalued. Since PBL relies on a constructivist framework and on collaborative learning, group dynamics play a critical role in the learning process (Savery & Duffy 2001) as well as during the assessment and evaluation phases of a course (Gijbels et al. 2005; Gijbels & Dochy 2005).

In the PBL learning environment, students from a large cohort are generally distributed in groups of 6 to 14 where course material is mastered through group-directed inquiry based around a problem or situation given to students by the tutor (Boud & Feletti 1997). We have implemented a biology program using a PBL format based on the practices of McMaster-Maastricht universities whereby our groups are typically 12 students that meet twice a week with a tutor (De Graaff & Kolmos 2003). We may have several tutors facilitating a given cohort all using the same problem. This presents challenges for educators tasked with providing similar educational experiences and assessment to these multiple groups of students. Students are tasked with identifying specific objectives, formulating and testing hypotheses using information such as primary literature and textbooks in an approach akin to the scientific method (Duch et al., 2001). During this process, tutors (instructors) are tasked with verifying that course objectives are covered during these discussions and intervene when necessary to clarify or redirect discussions. In our program, all students, regardless of group assignment, are evaluated with identical exams that are administered throughout the course. Under these circumstances significant variation in test scores often arises between groups suggesting that students are not receiving similar educational experiences among groups.

Potential sources for variation in test scores between groups may include differences in the quality/capacity of the tutor (Neville, 1999), differences in social dynamics between students within groups and even the initial selection and formation of groups prior to the course (Lowry et al. 2010; Lohman & Finkelstein, 2000). Selection and formation of groups differs fundamentally from other potential sources of between-group variation in that it occurs independently from the role of tutors or interaction among students within a group and its effects extend to all groups. Thus, variation in test scores due to group assignment may inescapably obscure the variation in tutors' capacity and student dynamics. As a consequence, differences in tutor quality/capacity determined by evaluations that are linked to student performance may be effectively masked unless differences among tutors are greater than variation created from group assignment. Likewise, the importance of within-group dynamics among students for test scores must exceed variation related to group assignment in order to be detectable. For these reasons, it is useful to quantify the variation in test scores between groups attributable strictly to group assignment as a baseline prior to the assessment of other sources of variation.

Relative to other factors, quantifying the effects of group assignment on the variability in test scores between groups is relatively straightforward. Each cohort of students enrolled in a course constitutes a random sample from larger population of students. If group assignment is based on random allocation of students to groups, the initial cohort of students is sampled without replacement. As with all sampling, allocation of students to groups and the variability of test scores between groups will depend on both the size and the number of groups. If the underlying distribution of test scores can be estimated for the larger student population, the allocation of students to groups can be analyzed via simulation to quantify the inherent variability related to the formation of groups.

Here we present simulations characterizing the variation in test scores between groups in a university-level course in a PBL program that is attributable exclusively to the initial selection and formation of groups. We compared the empirical variability between groups in a zoology course given between 2014-2017 based on tests given mid-term and at the end of the course to assess the relative variation attributable to other factors such as differences in quality/capacity of instructors or differences in group-dynamics. Assuming that maintaining similar educational experience between groups is desirable, we also suggest strategies for how a critical consideration of the allocation process of students into smaller groups could be changed to minimize between-group variation in a PBL context.

METHODS

We analyzed test scores from an introductory, university-level course in zoology in a PBL program in biology for our study. The course is required material for the baccalaureate program in biology at the Université du Québec à Montréal. The course counts for 7 credits of the total 90 credits within the program and takes place over 7.5 weeks during the second trimester. In this program, courses are offered sequentially rather than concurrently, thus students are involved only in this course during this period. During the second trimester, the course counts for more than half of the 13 credits offered. Enrolment in the course is ranges from ca. 75-85 students in a given year.

The course is divided into both practical and theoretical aspects. Practical aspects of the course are taught in a laboratory setting in larger groups where students concentrate on learning external and internal morphology and taxonomic identification. Theoretical aspects of the course are taught in smaller groups and led by a tutor (instructor) for the duration of the course. Each tutor is responsible for 1 (or less frequently 2) groups. Thus, in a given year there are typically 5-7 instructors assigned to different groups. Subject matter during this aspect of the course revolves around understanding phylogenies and major evolutionary transitions seen throughout the radiation of multicellular animals. Each class period, students are presented with

a short document that describes a problem/situation related to course subject matter. In these small groups with the aid of the tutor, students are expected to identify learning objectives and develop and use hypotheses to guide their inquiry. Once established, students verify these hypotheses using pertinent sources of information such as assigned reading from textbooks, current popular science writing in the media and peer-reviewed scientific literature. During this period, the tutor may intervene to clarify course material and assure that hypotheses and discussions revolve around material pertinent to the course. Students then meet to share and compare information responding to each hypothesis. The synthesis produced in these meetings serves as the base of course material that students use to prepare for exams. The tutor's role through this process is to guide students in the formulation of clear and verifiable hypotheses, to assure that pertinent subject matter is addressed with accurate information and to verify that students have met specific learning objectives for the course. For our analysis, we concentrated only on test scores from exams related to theoretical aspects of this course that were linked to performance in small groups with a single tutor.

Theoretical aspects of the course are evaluated twice during the course as a mid-term and final exam. Each exam has the same format and consists of 40% long-form essay response, 50% short response and 10% 'fill in the blank-type' responses and is based on a total 100 points. Long-form essay questions are broad questions that require students to integrate material from several problems/situations to support their responses that can be up to both sides of a single page. Short response questions require less development and target more specific aspects of the course. Fill-in the blank type responses consist of students correctly naming organisms associated with a phylogenetic tree that emphasizes evolutionary relationships among taxa. Exam scores (based on 100 points total) are then given letter grades as follows: A+>88%, A 85-87%, A- 82-85%, B+ 78-82%, B 75-78%, B- 72-75%, C+ 70-72%, C 68-70%, C- 65-68%, D+ 63-65%, D 60-63%, E<60%.

STATISTICAL ANALYSIS

For our analysis, we compiled mid-term and final exam scores between 2014 and 2017. We combined scores from all years to provide an overall empirical distribution of test scores for each exam. The empirical distribution was then used to choose reasonable parameters for beta distributions that were used in simulations. Beta distributions are convenient representations for student grades because 1) they are bounded between 0 and 1, which can be easily translated to 0 to 100% a standard scale for evaluations and 2) they can be negative skewed which captures well the large range of values associated with a failing grade (usually all notes <60). Beta distributions depend on two shape parameters (α and β). Beta distributions with parameter values of α between 4 and 5 and β between 2 and 2.5 have modes near 70% and a strong negative skew similar to the distributions of grades observed in university courses. For our analysis, we used four beta distributions where α=4 or 5 and β=2 or 2.5 to provide a realistic range of possible

student populations. We then simulated 1000 virtual cohorts where 80 values representing students were drawn at random from a beta distribution and then separated into 8 groups of 10. For each cohort, we calculated the maximum difference in mean test scores between groups. This value characterizes the largest difference in the mean test scores between groups in any simulated cohort. We summarized the differences generated through simulation to quantify the extent to which the selection/formation process influences between group variability in test scores. These simulations represent the between group variation in test scores that occurs exclusively as a product of random sampling without replacement and group assignment.

We also compared to the maximum difference in mean test scores between groups generated from randomizations of empirical data for mid-term and final exams. For these randomizations, student test scores were randomized and assigned without replacement into individual groups where each group had a minimum of 10 students. The number of midterm and final test-scores used in these simulations differed slightly within each year as some students dropped the course between the mid-term and final exam. Based on 1000 randomizations and group assignments, we estimated the distribution of the maximum difference in mean test scores between groups. These simulations represent the between-group variation in test scores that would occur following the selection and allocation of students to groups as well as other factors such as differences in quality/capacity of instructors or differences in group dynamics.

We then compared between group differences in test scores between both sets of simulations to determine the extent to which factors other than the selection process and group formation contribute to differences in test scores between groups. We hypothesized that on average the differences in test scores between groups derived from empirical simulations would be greater than differences in test scores derived from beta distributions because of additional factors including differences among tutors and group dynamics. All randomization and simulations were made using the sample function in RStudio (RStudio Team 2015).

RESULTS

The combined mid-term and final test scores from 2014-2017 both had modes at 70%, negative skew and a large proportion of observed values within 60-80% (Fig 1). Overall variability around the mode decreased between the mid-term and final exams with fewer scores lower than 60% observed in the final exam than in the mid-term exam. Empirical distributions were similar qualitatively to beta distributions with shape parameters ranging from 4-5 and 2-2.5 for α and β respectively (Fig 2). Beta distributions with α and β ranging from 4-5 and 2-2.5 respectively maintained modes at or near 70%, negative skew and a large proportion of values falling within 60-80%.

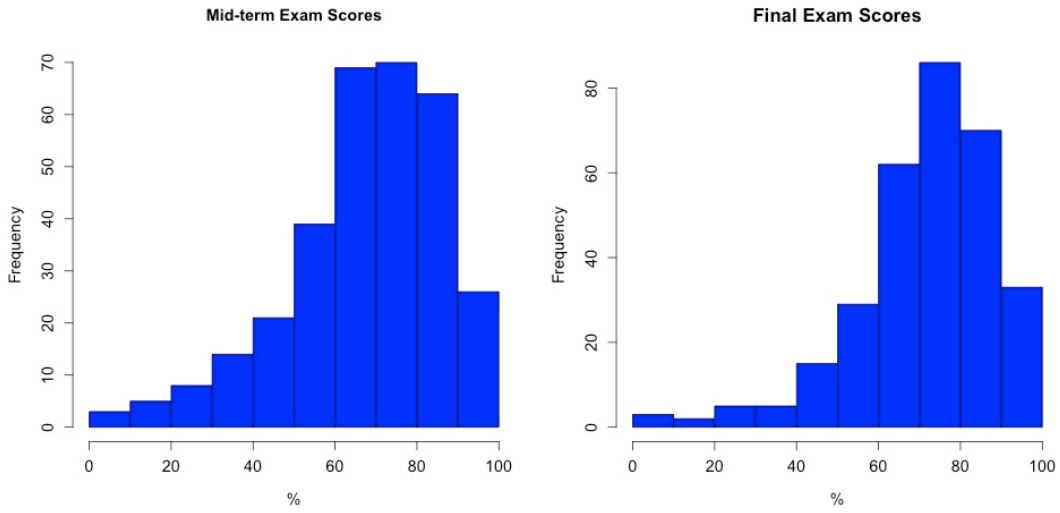

Figure 1. Empirical distribution of (left) mid-term and (right) final exam scores between 2014 and 2017 from introductory course in zoology taught through problem-based learning.

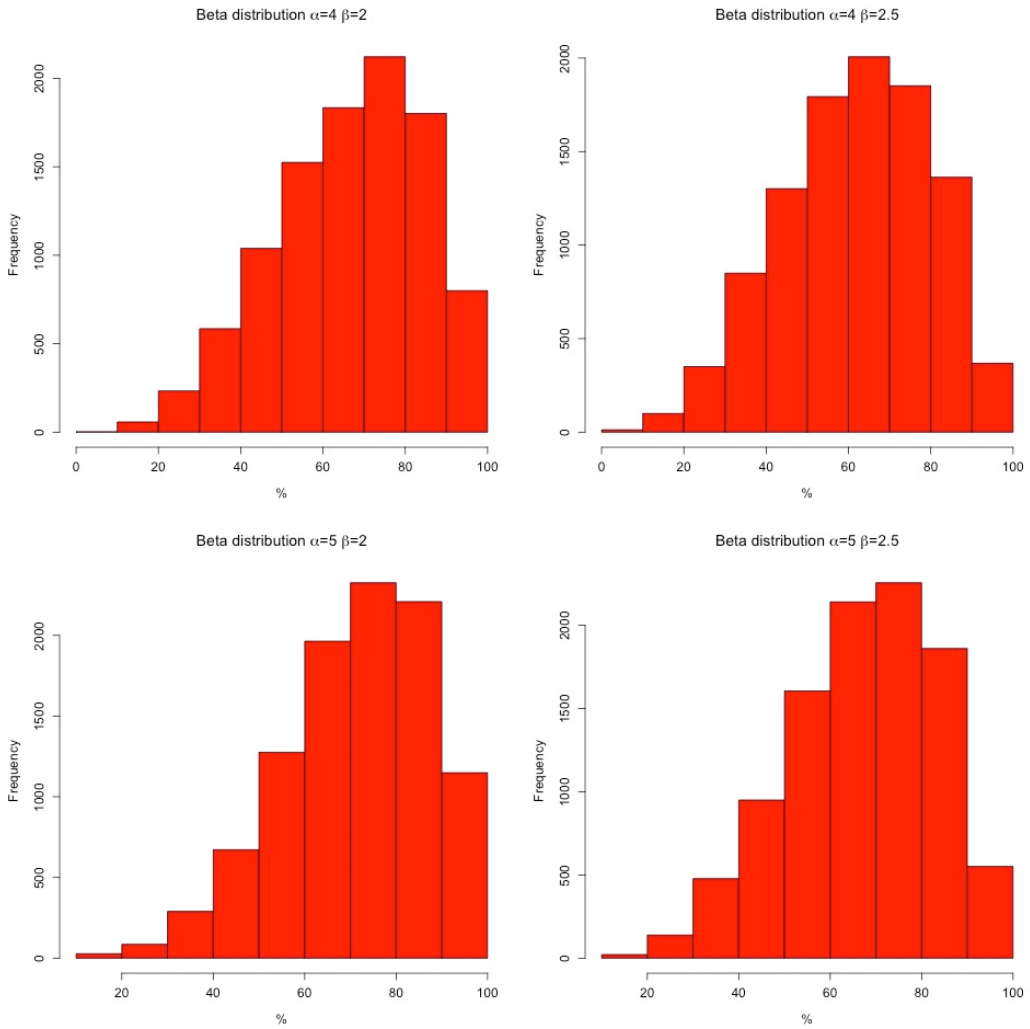

Figure 2. Histograms of 10,000 observations drawn from beta distributions with shape parameters (α, β) ranging from α=4 or 5 and β= 2 or 2.5.

Given the range of shape parameters used in our simulations, we found that average value for the maximum difference in mean test scores between groups ranged between 14.4 and 16.2 points (Fig 3). The majority of values (25-75% quartiles) for the between-group differences in test scores drawn from beta distributions fell between 11.5 and 19.2 points. Extreme differences between group means (>1.5 * the interquartile range which corresponds to a maximum difference of >25 points between groups) accounted for only a small proportion (ca. 1%) of the simulation values observed under any specific parameter combination. Extreme differences between group means were slightly larger in simulations where the β parameter of the underlying distribution equalled 2 (13 and 12 extreme value differences between groups when α=4 and 5 respectively) than when the β parameter equalled 2.5 (7 and 9 extreme value differences between groups when α=4 and 5 respectively).

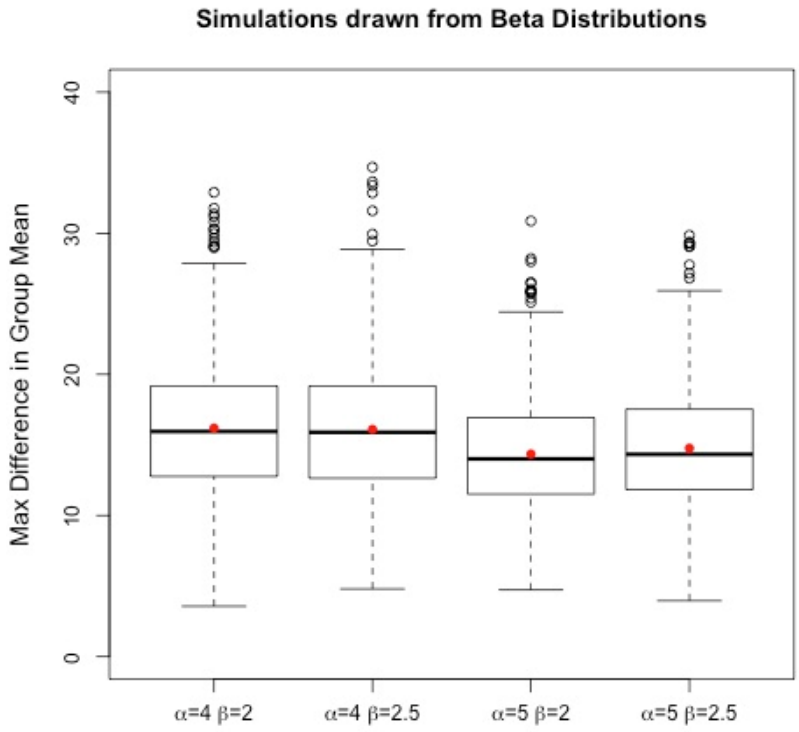

Figure 3. The maximum difference in mean test scores between problem-based learning groups from 1000 simulated cohorts of students. Test scores were drawn from four beta distributions (with shape parameters α=4 or 5 and β=2 or 2.5) and randomly assigned to learning groups. Box plots depict median values (solid black line), 25% and 75% quartile (boxes) and 1.5 times the interquartile range (whiskers). Red dots depict the mean.

When actual test scores were randomized and re-allocated to groups, mean values for the maximum difference in test scores between groups ranged from 12.2 to 18.2 for mid-term exams and from 12.5 to 17.3 for final exams (Fig 4). The overall range of between-group differences in test scores was also similar for both mid-term and final exams over the four years

examined in this study. Interquartile ranges (25-75% quartiles) for between-group differences in empirical test scores over all four years ranged between 9.6 and 21.4 for the midterm exam and 10.1 and 20.2 for the final exam (Fig 4). However, the inter-annual patterns in between-group differences of empirical test scores was different between the mid-term and final exams. Between-group differences in mid-term test scores were greater in 2015 and 2017 than in 2014 and 2016, while between-group differences in final exam test scores were similar in 2014-2016 but increased in 2017 (Fig 4). As with simulations derived from beta distributions, extreme values in simulations based on empirical data (>1.5 * the interquartile range) accounted for only a small proportion (ca. 1%) of the differences observed in any given year. However, in simulations based on empirical data, we observed 4 extreme values (out of 4000) where between-group variation in test scores was extremely low (<5 points) (Fig 4).

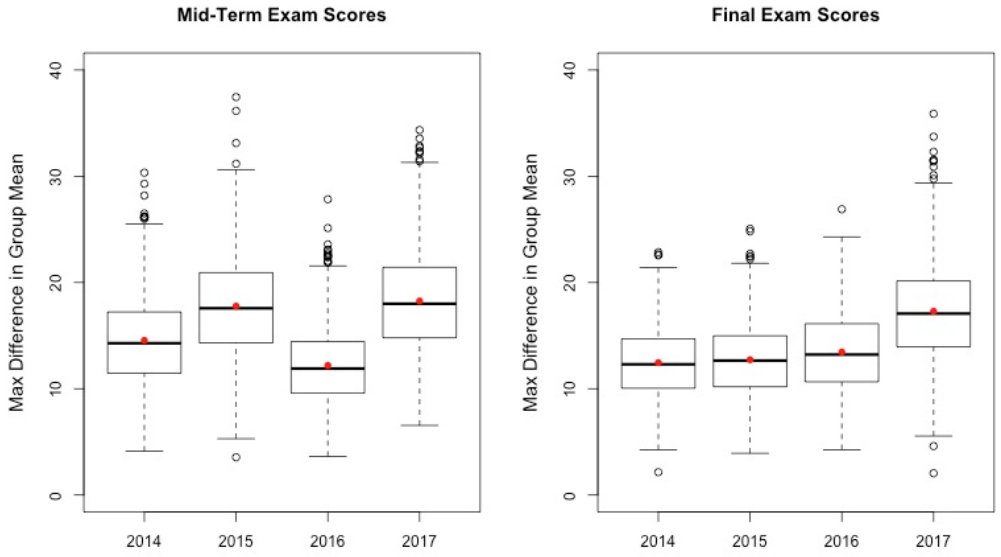

Figure 4. The maximum difference in mean test scores after randomization and reallocation of empirical mid-term and final test scores among groups between 2014 and 2017. Box plots depict median values (solid black line), 25% and 75% quartile (boxes) and 1.5 times the interquartile range (whiskers). Red dots depict the mean.

When we compared simulations drawn from empirical data with those drawn from beta distributions, between-group differences in empirical test scores only marginally exceeded those drawn from beta distributions and not in every year. The beta distribution with parameters $\alpha=5$ and $\beta=2$, had smallest between-group differences in test scores thus provided an analytically conservative benchmark for comparisons. The average difference between-groups in empirical test scores for mid-term exceeded simulations based on this conservative beta distribution by no more than 4 points in any given year and were less than simulation results in 2016. Likewise, the average difference between-groups in empirical test scores for final exams exceeded simulations from the conservative beta distribution by less than 3 points and only in

2017. In preceding years, the average difference between-groups in empirical test scores for final exams never exceeded the conservative beta distribution. Differences between simulations drawn from empirical data and those based on the other beta distributions used in our study will necessarily be less, further stressing the relative importance of the selection and group formation process on between-group differences in test scores.

DISCUSSIONS

Regardless of the parameters chosen for our simulations, significant intergroup variation of between 14.4-16.2 points was imparted solely through random assignment of students to groups. For students, this variation corresponds to the difference of 2 letter grades in typical grading scales where letter grades are separated by 10 points. For tutors, between-group variation from random assignment obscures potential differences between tutors unless differences in tutors' performance between groups exceed ca. 14-16 points. Likewise, quantitative evaluations of social dynamics within groups including student performance attributable to group size will be hampered unless it exceeds the variation inherent in the assignment of students to groups.

Incorrectly attributing between-group variation caused by group assignment to other sources of variation such as tutor performance can cause significant loss of time and result in ineffective evaluation strategies. It is our experience that tutors themselves often attribute between-group variation in test scores to differences in severity or generosity of different tutors during the correction of exams. However, it has been reported that tutors may over-rate students of the group reflecting a bonding affect (Whitfield & Xie, 2002, Cohen et al. 1993). This often leads to protracted but pointless discussions related to correction strategies. Common discussions revolve around the questions whether 'it is better to standardize corrections by having one tutor correct exams (or individual exam questions) from all groups' or 'should extremely detailed correction guides be prepared and rigidly implemented' to assure homogeneity between groups during the correction phase. Both strategies could decrease between-group variation, however these effects would be minimal. Our results suggest that after removing the between-group variation attributable to formation of groups, which under a conservative scenario based on a beta distribution with parameters $\alpha=5$ and $\beta=2$ would account for a differential of 14 points between groups, implementing additional ad-hoc correction strategies to minimize between group variation could -at best- reduce between group differences up to 4 points. In our study such minor reductions would be possible in only 3 of the 8 evaluations (2015 and 2017 mid-terms and 2017 final exam). For the remaining five evaluations the empirical between-group differences in test scores falls with the range of variation attributable solely to the group formation process.

There are additional costs associated with ad-hoc correction strategies. Strategies by which individual tutors correct exams or individual exam questions across groups do little to correct generosity or severity during the correction phase and only shift these disparities to differences in student performance on individual questions. Highly detailed correction guides often do not capture creative/unforeseen aspects in student responses-particularly in long-form essay responses to more open-synthetic type questions. It is not our position that clear correction guides and similar expectations among tutors are unwarranted during the correction phase. It is our position that these strategies will have negligible effects on between-group variation in exam scores.

One alternative is to consider a non-random assignment of students to groups. A simplistic assignment strategy would allocate students to groups based on prior performance in PBL courses. For introductory courses such as the example used in this study, prior evaluations made in PBL courses may be limited or unavailable. In cases where prior evaluations are available, there may be the practical limitation of an insufficient number of students with elevated performance that can be dispersed among groups. This has profound implications on learning strategies and group dynamics of students within a group and has been discussed in the context of behavioural ecology as the 'producer-scrounger' argument (Vickery 2013). This argument is premised on the idea that a limited number of students who contribute to group discussion during the problem or situation given to the students (ie. 'producers') promote an inverse effect by which other students contribute less (ie. 'scroungers') choosing to profit disproportionately from 'producers'. One prediction of this argument is that there is an optimal number of level of scrounging. Thus, as the number of producers in a group increases, the number of scroungers decreases. However, below the optima predicted under the 'producer-scrounger' model, scroungers should take on a greater role in class discussions and thus become producers. While this argument has been founded on ecological principals, overall performance of a group likely involves additional social dynamics among students within the group (shaming, competition, the development of cohesion and cooperation within a group). While it is intriguing to think that there may be some 'magic' formula for group assignment and group size in PBL learning continues to be an active area of research, any kind of non-random assignment can also be strongly criticized as favouring/disfavouring students depending on the criteria chosen. This highlights the need for demonstrable indicators of student performance if non-random group assignment strategies are to be adopted. If such indicators do not exist and assignment criteria cannot be justified, then tutors should learn to live with significant between-group variation in PBL courses.

For tutors and directors of PBL teaching units, the relatively large between-group variance that arrives from the group formation process presents a challenge for evaluating tutor performance. Reliance on student test scores as a metric of tutor capacity is clearly limited by the variation introduced through the formation of groups. Such strategies could only see extremes in tutor performance and would still rely on the untested assumption that variation between groups does

not occur from other sources such as group dynamics. In this context, we suggest that other metrics of tutor performance such as thoughtful evaluations made by students may be more useful.

ACKNOWLEDGEMENTS

We appreciated conversations with Jonathan Lachance (UQAM) and thoughtful commentary provided on initial versions of this manuscript by William Vickery (UQAM) and Kahlid Addi (Université de la Réunion).

References

Akcay, B. (2009). Problem-based learning in science education. *Journal of Turkish Science Education*, 6:26-36.

Azer, S.A. (2003) Assessment in a Problem-based Learning Course. Biochemistry and Molecular Biology Education 31(6): 428-434.

Barrows, H. S. (1996). Problem-based learning in medicine and beyond: A brief overview. *New Directions for Teaching and Learning*, (68): 3–12.

Bell, S. (2010). Project-Based Learning for the 21st Century: Skills for the Future. *The Clearing House: a Journal of Educational Strategies, Issues and Ideas*, *83*(2): 39–43.

Boud, D & Feletti. G. (1997). The Challenge of Problem-based Learning. London: Kogan Page Publishing.

Cohen, G.S., Blumberg, P., Ryan, N.C. & Sullivan, P.L. (1993) Do final grades reflect written qualitative evaluations of student performance? Teaching and Learning in Medicine 5: 10–15.

Decuyper, S., Dochy, F., & Van den Bossche, P. (2010). Grasping the dynamic complexity of team learning: An integrative model for effective team learning in organizations. *Educational Research Review*, 5:111-133.

De Graaff, E. & A. Kolmos. 2003. Characteristics of Problem-Based Learning. Int. J. Engng Ed. Vol 19, No. 5 : 657-662.

Duch B, Groh SE, & Allen DE. (2001). The Power of Problem-Based Learning: a Practical "How To" for teaching undergraduate courses in any discipline. Sterling, VA: Stylus Publishing.

Galand, B., & Frenay, M. (2005). L'approche par Problèmes et par Projets dans l'Enseignement Supérieur: Impact, Enjeux et Defies, Louvain-la- Neuve: Presses Universitaires de Louvain.

Gijbels, D., & Dochy, F. (2005). Effects of problem-based learning: A meta-analysis from the angle of assessment. *Review of Educational Research*, 75:27-61.

Gijbels, D., Van de Watering, G., Dochy, F. & Van den Bossche, P. (2005). The relationship between students' approaches to learning and the assessment of learning outcomes. *European Journal of Psychology of Education*, 4:327-341.

Heycox K, & Bolzan N. (1991). Applying Problem-Based Learning in First-Year Social Work. In D. Boud & G. Feletti (Eds). The challenge of problem based learning. (pp. 186-193). London, Kogan Page Press.

Lohman, M. C., & Finkelstein, M. (2000). Designing groups in problem-based learning to promote problem-solving skill and self-directedness. *Instructional Science*. 28:291-307.

Lowry, P. B., Zhang, D., Zhou, L., & Fu, X. (2010). Effects of culture, social presence, and group composition on trust in technology-supported decision-making groups. *Information Systems Journal*, 20:297-315.

Macdonald, R. (2005) Assessment strategies for enquiry and problem-based learning. In *Handbook of Enquiry & Problem Based Learning*. Barrett, T., Mac Labhrainn, I., Fallon, H. (Eds). Galway: CELT, 2005. Released under Creative Commons licence. Attribution Non-Commercial 2.0. Some rights reserved. http://www.nuigalway.ie/celt/pblbook/

Macdonald, R.F. & Savin-Baden, M. (2004) "A Briefing on Assessment in Problem-based Learning," *LTSN Generic Centre Assessment Series*. Available on the Higher Education Academy's Resource Database at: www.heacademy.ac.uk/resources.asp?process=full_record&section=generic&id=349

Mills, J.E. & Treagust, D.F. (2003). Engineering education – Is problem-based or project-based learning the answer? Australasian journal of engineering education, 3(2): 2-16.

Neville, A. J. (1999). The problem-based learning tutor: Teacher? Facilitator? Evaluator? *Medical Teacher*, 21:393-401.

Prince, M. (2004). Does active learning work? A review of the research. *Journal of Engineering Education*. 93(3):223-231.

Prince, M.J. & Felder, R.M. (2006). Inductive teaching and learning methods: definitions, comparisons, and research bases. Journal of Engineering Education, 95(2): 123-138.

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/

Savery, J. R., & Duffy, T. M. (1995). Problem based learning: An instructional model and its constructivist framework. *Educational Technology*. CRTL Technical Report 16-01

Savin-Baden, M. (2004). Understanding the impact of assessment on students in problem-based learning. Innovations in Education and Teaching International. 41:221-233.

Strobel, J., & Van Barneveld, A. (2009). When is PBL more effective? A meta-synthesis of meta-analyses comparing PBL to conventional classrooms. *Interdisciplinary Journal of Problem-Based Leraning* 3:44-58.

Van den Bossche, P., & Gijselaers, W. H. (2006). Social and cognitive factors driving teamwork in collaborative learning environments: Team learning beliefs and behaviors. *Small Group Research*, 37(5): 490-521.

Vickery, W. L. (2013). Producing and scrounging during Problem Based Learning. *Journal of Problem Based Learning in Higher Education*, *1*(1): 36–52.

Whitfield, C.F. & Xie, S.X. (2002) Correlation of Problem-Based Learning Facilitators' Scores with Student Performance on Written Exams. Advances in Health Sciences Education 7:41-51.