Reliability in the Assessment of Program Quality by Teaching Assistants During Code Reviews

Michael James Scott
Department of Computer Science
Brunel University London
United Kingdom
michael.scott@brunel.ac.uk

Gheorghita Ghinea
Department of Computer Science
Brunel University
United Kingdom
george.ghinea@brunel.ac.uk

ABSTRACT

It is of paramount importance that formative feedback is meaningful in order to drive student learning. Achieving this, however, relies upon a clear and constructively aligned model of quality being applied consistently across submissions. This poster presentation raises concerns about the inter-rater reliability of code reviews conducted by teaching assistants in the absence of such a model. Five teaching assistants each reviewed 12 purposely selected programs submitted by introductory programming students. An analysis of their reliability revealed that while teaching assistants were self-consistent, they each assessed code quality in different ways. This suggests a need for standard models of program quality and rubrics, alongside supporting technology, to be used during code reviews to improve the reliability of formative feedback.

Categories and Subject Descriptors

K.3.2 [**Computers and Education**]: Computer and Information Science Education.

Keywords

Programming, Code Review, Grading, Quality, Assessment, Reliability, Concordance, Agreement, Consistency.

1. INTRODUCTION

Guidance is important when first learning computer programming. This is because students often need help to develop an appreciation for program quality. Such guidance often consists of formative feedback provided during code reviews. However, in large undergraduate cohorts, such code reviews may be conducted by teams of teaching assistants.

For feedback to be meaningful to students, it should be clear, reliable and constructively align with relevant learning objectives (c.f. [3, 5]). This is because conflicting feedback from different sources could cause confusion. Previous work suggests that reviews by experienced faculty tend to be correlated, but different reasoning is sometimes applied [1]. It is not clear, then, whether assessments made by teaching assistants would be as consistent. Of particular concern is that assessments of program quality may reflect more on the reviewer than on the student (see [4] for detail on the idiosyncratic rater effect).

ITiCSE'15, July 6–8, 2015, Vilnius, Lithuania.
ACM 978-1-4503-2078-8/13/07.

Table 1: Reliability of Assessments ($\alpha >= 0.667$)

Measure	Reliability α
Self-Consistency	.841
Agreement Between Teaching Assistants	.607
Agreement with Faculty Assessments	.522

2. FINDINGS

Five teaching assistants, each with at least one year of experience, reviewed 12 purposely selected programs submitted by first-year computing students and made holistic assessments of their quality using a 3-point scale (pass, merit, distinction). Minimal instruction was provided to reflect a less formal formative (rather than summative) context. After two weeks, they re-reviewed the programs. On each occasion the programs were presented in a random order and some elements (e.g., identifiers) were transformed. The data were analysed using Krippendorf's alpha [2].

The results, shown in Table 1, show that while the assessments were adequately self-consistent, there was low inter-rater reliability and there was considerable disagreement with ratings provided by faculty. This finding suggests that teaching assistants use different notions or standards of program quality when conducting code reviews and therefore need support. As such, this study provides a foundation for future work on the development and evaluation of code review processes, program quality rubrics, and supporting technologies.

3. REFERENCES

[1] S. Fitzgerald, B. Hanks, R. Lister, R. McCauley, and L. Murphy. What are we thinking when we grade programs? In *SIGCSE '13*, pages 471–476. ACM, 2013.

[2] A. Hayes and K. Krippendorff. Answering the call for a standard reliability measure for coding data. *Commun. Methods & Measures*, 1(1):77–89, 2007.

[3] A. Pears, J. Harland, M. Hamilton, and R. Hadgraft. Four feed-forward principles enhance students' perception of feedback as meaningful. In *LaTiCE '14*, pages 272–277. IEEE, 2014.

[4] S. E. Scullen, M. K. Mount, and M. Goff. Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, 85(6):956, 2000.

[5] M. Stegeman, E. Barendsen, and S. Smetsers. Towards an empirically validated model for assessment of code quality. In *Koli Calling '14*, pages 99–108. ACM, 2014.