# What Can We Learn From Quantitative Teaching Assistant Evaluations?

Elizabeth Patitsas
University of Toronto
40 St George St.
Toronto ON M5S 2E4
patitsas@cs.toronto.edu

Patrice Belleville
University of British Columbia
2366 Main Mall
Vancouver BC V6T 1Z4
patrice@cs.ubc.ca

## ABSTRACT

The teaching assistant plays an important role in teaching computer science at large research-intensive universities. We conducted an exploratory study examining what quantitative teaching evaluations can tell us about our TAs. We found that evaluations were highly coarse-grained: students did not differentiate between the different criteria in the evaluations, and that TAs working in pairs were evaluated as one unit. We found that poor TA evaluations were negatively correlated to student retention. We also found that TAs teaching more than three lab sections a week had lower evaluations. Finally, we found no correlation between TA evaluations and TA experience, students' final grades, and TA gender. We note that quantitative evaluations paint an incomplete picture of TAs' performance, and that more work is needed to provide TAs with formative assessment.

## Categories and Subject Descriptors

K.3.2 [**Computers and Information Science Education**]: Pedagogy, education research

## General Terms

Measurement

## Keywords

Computer education, teaching assistants, teaching evaluations, retention

## 1. INTRODUCTION

Computer science is a young discipline; we still are learning not only how to effectively teach the material, but also how to effectively support the teachers. While work has been done on supporting K-12 teachers (e.g. [8]), little has been done for teaching assistants (TAs). At the large research-intensive universities in North America, TAs make up the majority of the teaching staff [1]. At our institution, TAs' duties include teaching labs, tutorials, grading, and office hours; senior TAs may also develop curriculum, and manage and train other TAs.

Previous work in CS education has found that student access to TAs contributes to success in introductory computer science [17], particularly in large CS1 courses at research-intensive universities [14]. Effectively training and supporting TAs has been identified as important to making CS more minority-friendly [15]. Providing TAs with student evaluations has been identified as one of many practices that is important for improving TA quality [15].

At the University of British Columbia, we hire about a hundred TAs a year – compared to 55 faculty members. These TAs provide 46% of the contact hours in our first and second-year courses. A majority of these TAs are hired as part of MSc/PhD guaranteed funding packages – and not for their teaching ability.

As educators, we wish to provide our computer science students with the best possible teaching we can. For large research institutions, this means that training and supporting teaching assistants is vital. So what can we do to help our computer science TAs?

We know from the education literature that feedback is important for novice teachers [16]. From our own previous work on computer science TAs, we know that teaching evaluations provide a source of feedback that TAs use to improve their teaching [11]. Indeed, at our institution, this is the only formal source of feedback that TAs receive. So what do these teaching evaluations actually tell us?

In this study, we examined the quantitative portion of the TA evaluations used at our institution. Our goal was to identify what, if anything, these evaluations can be linked to in terms of the TAs' characteristics (experience, workload, etc), performance, or student success (grades, retention). This was a preliminary study, the first step in a larger goal of providing better formative assessment to teaching assistants.

### 1.1 Background information

There has been a large amount of work in the education and psychology literature finding that teaching evaluations don't measure student learning, and are confounded by a myriad of factors [4]. For example, a recent study of reviews on RateMyProfessors.com found that the number of chili peppers (a rating of the professor's attractiveness) is correlated to teaching scores [3]. Another study [7] found that teaching evaluations actually *negatively* correlate to


*WCCCE'12,* May 4–5, 2012, Vancouver, British Columbia, Canada.

how much students learn. Teaching evaluations, instead, appear to be more of an evaluation of how much the students *like* their teacher [7].

While these evaluations may not measure student learning, this is not to say they measure nothing. One may expect that how much students like their teachers can be linked to students' engagement with course material and how much more computer science they take afterwards. After all, having role models has long been identified as good for students, particularly those of minority groups [20].

Teaching assistants can be role models to our students [10] – and previous work has found that they have an influence on student retention [9] and even their students' final grades [12][1]. The TA hence has a significant, and more "hands-on" role in their students' education. Yet, research on TAs is described in the literature as "under-developed" [6].

Muzaka's 2009 study [6] identified three aspects about TAs: TAs' subject knowledge – a weakness of TAs; the relative informality and flexibility of TAs' teaching styles – a strength of TAs; the relative approachability of TAs and their ability to relate to their students – another strength of TAs. Related to the informality and approachability of TAs is the smaller age gap between them and their students. Worthington [19] found that teaching evaluations are correlated to age – the approachability that a small age gap affords hence leads us to expect TAs to be well-rated by their students. And the link to student retention matters even more to us given the low student retention we have in CS.

Receiving positive teaching evaluations improves TAs' self-confidence – which in turn improves their teaching [1]. Independent of how much their students are learning, TAs want to be well-liked [1], reinforcing the TA evaluation as important to this population of computer science teachers. Teaching evaluations remain commonly used in evaluating job candidates [16] and teaching awards – indeed, the quantitative portion of TA evaluations is what is used to give out TA awards in our computer science department.

## 2. METHODS

Our approach in this study is exploratory, using quantitative methods. We began by performing a literature review to identify any possible, quantifiable characteristics that may be linked to TA evaluations. We then proceeded to test whether each of these characteristics can be statistically linked to TA evaluations.

Anonymized student records were acquired for students enrolled in CS1 between 2006 and 2009, and in LOGIC1[2] and CS2 between 2007 and 2009. These records contained: their lab, lecture and tutorial section codes; their degree programme; and whether they have since taken the next four CS courses. There were records for 2773 students who had taken CS1, 1025 in LOGIC1, and 1781 in CS2.

Secondly, we acquired anonymized TA evaluation records for the TAs of those courses during the same time-frame. The records contained the codes of the sections that the TAs taught, whether the section was a lab or tutorial, the contract the TA was hired with, and their gender. It also contained their results to the end of term TA evaluation with regards to each of the provided criteria; along with the number of respondents, and the standard deviations. We acquired records for 355 evaluations: 194 for CS1, 79 for LOGIC1, and 82 for CS2.

For the purpose of this study, we will consider a relationship to be statistically significant if $p < 0.001$. We choose to use a value less than the standard 0.05 since we are performing dozens of statistical tests in this study, and wish to reduce the likelihood of false positives.

[1]Of course, faculty can also influence these things; for example de Paola [2] found a link between instructor's teaching scores and student retention.

[2]LOGIC1 is a first-year course in proof techniques, discrete mathematics and digital logic, required for our CS programme.

**Table 1: Pearson Correlation values for CS1 TA criteria, n = 194. All R values have $p < 0.0000001$**

| | W.P. | Help. | Consid. | E.U. | E.I. |
|---|---|---|---|---|---|
| Well prepared | | | | | |
| Helpful | 0.8 | | | | |
| Considerate | 0.7 | 0.8 | | | |
| Easily understood | 0.7 | 0.8 | 0.6 | | |
| Effective instr. | 0.8 | 0.9 | 0.8 | 0.9 | |

## 3. TA EVALUATION CRITERIA

TAs at the University of British Columbia are evaluated with five standard criteria. We found that these five criteria used to evaluate TAs are all strongly correlated. The five criteria are evaluated on a 5-point Likert scale, and are:

- Well prepared (W.P.)
- Helpful (Help.)
- Considerate of students (Consid.)
- Easily understood (E.U.)
- An effective instructor (E.I.)

As we see in Table 1, these criteria are all *each* significantly and strongly correlated. It is unclear from these data whether this is an artifact of the particular criteria used, or whether students do not discriminate between any criteria when providing evaluations. Either way, we are given the impression that the different criteria do not provide a fine level of detail about how the TA is perceived by their students. We speculate that these quantitative evaluations may have little use to the TA in judging their performance beyond a very coarse-grained picture. Qualitative TA evaluations may prove more useful and able to identify strengths and weaknesses of a given TA; examining them is an area of possible future work.

## 4. TA CHARACTERISTICS

### 4.1 Gender

We found no differences between how female and male TAs were ranked by their students. Of the five TA evaluation criteria in each of the three courses, there was no statistically significant difference between the genders (recall $p < 0.001$). There was one exception: in CS2, the female TAs were rated better with regard to understandability than the males.

While there were no real gender-based differences in TA evaluations, there was a difference in proportions. For example, the undergraduate student body in the courses studied is 34% female – but 44% of the undergraduate TAs are female. Overall, however, 26% of all the TAs are female – the graduate TAs are disproportionately male.

## 4.2 Undergraduate TAs vs. Graduate TAs

We found that undergraduate TAs (UTAs) were ranked statistically significantly higher than graduate TAs (GTAs). Of the graduate TAs, those who had been hired under the hourly contract were rated better than those hired under the monthly contract. It should be noted that there are large differences between the GTAs hired under those two contracts. Monthly GTAs are mostly first-year MSc students; they receive a monthly salary that is guaranteed to them as part of their funding package. Hourly GTAs, however, are mostly experienced graduate students; these TAships are not guaranteed and generally only given to experienced GTAs.

In CS1, the UTAs performed best, with an average effectiveness score of 4.5 out of 5. It is somewhat less in LOGIC1, at 4.4, and lesser still in CS2, at 4.2. It would appear that UTAs perform best in the earlier computer science courses. We discuss this in more detail in the Discussion (Section 6.)

Indeed, by CS2, the hourly GTAs outperformed the UTAs. Across the board with each category, the UTAs and hourly GTAs had high average scores (above 4.0).

The monthly GTAs consistently were ranked lower than the other two types of TAs. In CS1 they are on par with the other two categories of TAs, at 4.4. But for LOGIC1 and CS2, they were ranked at 4.0 and 3.9. Understandability (rating of "Easily understood") was where they performed the worst, and was most pronounced in CS2, where their average score was a 3.7. This is probably a result of the population involved: many of these TAs are international students with less English-speaking experience than the UTAs or hourly GTAs. In CS2, there are more TA pairs where both TAs are GTAs, which is perhaps responsible for the effect being strongest in this course

The monthly GTAs also had the highest variance in scores across the board – not surprising given that they are not hired for their proven ability to teach.

## 4.3 Experience

We found no statistically significant link between course-specific experience and TA evaluations. There are a number of reasons why this could be. One is that quantitative TA evaluations measure too coarsely to pick up any effect of experience. Another is that while more experienced TAs might be expected to do a better job as a TA, they would also be more confident – and hence more likely to enforce rules that may make them unpopular with the students. Indeed, other research on TAs has found that more experienced TAs describe their approach to teaching as being more "stern" than when they started out [11].

While there is evidence in the literature that TAs improve as teachers over time [13], this may not translate into improved student evaluations. For example, as more experienced TAs are "sterner" in enforcing rules [11] this may have a negative effect on TA evaluations which conflicts with any positive effect of experience.

## 4.4 TA Workload

We found a negative correlation between how many sections a TA taught and their TA evaluations. Specifically, we found this correlation in how helpful, easily understood, and effective they were ranked. This may indicate that overworked TAs are less effective, or at least less helpful and understandable.

Examining the data in more detail, we found that TA evaluation scores were similar between one-section and two-section TAs; three-section TAs performed worse, and four-section TAs – the maximum number of sections seen in our data – performed the worst of them all. It would appear that two sections is the best balance of maximizing a TAs' hours and keeping their evaluations high. It would also appear that four sections is more workload than a TA can comfortably handle.

It should be noted that TAs at our institution typically work 120 hours in a 12-week term. TA duties generally include teaching labs or tutorials, office hours, assignment grading, exam invigilation and exam grading. A lab section is typically 2-3 hours a week; in terms of weekly hours it appears that more than six hours of contact a week is the inflection point.

## 4.5 Pairing

We found that paired TAs received similar evaluations. Lab sections at the University of British Columbia are usually taught by pairs of TAs. These pairs were identified and the average effectiveness was computed for both. We began by looking at the distributions of the pairs' scores with regard to the different criteria. First, we compared these distributions to how the TAs performed individually. We noticed differences, and wished to probe this further. So, we shuffled the pairings randomly, and compared the real distributions to the random ones.

The randomized-pair distributions appeared as what we'd expect if TA scores were independent. However, the actual pair distributions looked different. The distribution becomes narrower and shifts upwards – in short, the evaluations are higher.

Furthermore, in the three courses, there were no TA pairs with an average effectiveness score less than 3.5, although there were a number of individual TAs who had such scores. In short: there were no pairs where *both* TAs were rated poorly. But is it the case that two "bad" TAs together bring each other up? Does being paired with a "bad" TA result in higher ratings for the other TA?

To examine this, we looked at the absolute value of the difference between the scores of TAs in a pair, and again compared this to the randomized pairs. What we saw was a noticeable downwards shift. TAs in pairs had less of a difference between their scores than if they were evaluated independently. It appears that TAs are evaluated as a team – for a typical poorly ranked TA, their partner tends not to be ranked very well either. A cynical interpretation of these data may be that students are failing to differentiate their TAs – perhaps they evaluate based on a setting-by-setting basis than teacher-by-teacher basis.

# 5. EVALUATIONS AND STUDENTS

## 5.1 Grades

We found absolutely no link between students' final grades and their TAs' evaluations. This does not discount the likelihood that a TA has an effect on their students' performance; indeed, the literature has found this to be the case [12]. It does, however, identify that whatever effect a TA has on their students is *not* captured by these quantitative evaluations.

## 5.2 Programme-Level Retention

We found that highly-ranked TAs' evaluations were uncorrelated to how many more computer science classes their students took. However, we did find a weak, negative correlation between low-ranked TAs' evaluations and how many more computer science classes their students took.

We quantified student retention by "retention score": the number of courses they took afterwards; e.g. for CS1, a student's retention score would be how many of the postrequisite first and second year courses they had taken. This score was then averaged for entire lab sections and was correlated to their TAs' evaluation scores for that section.

Analyzing all the TA evaluations together, there is no significant correlation between retention score and TA evaluations. One thing to note was that the evaluations fell into two clusters: above averages of 4.6, and averages between 4.0 and 4.5. A $k$-means cluster test confirmed that these clusters were indeed distinct.

For the upper cluster, we see no correlation between TA evaluations and retention score. For the lower cluster, a mild, negative correlation was found. It would appear that either students who are already unlikely to take more computer science are less keen about their TAs – or that "bad" TAs influence their students to take less computer science. Identifying which case is most likely is an area of possible future work.

# 6. DISCUSSION

While in this study we present some original findings, much of the work here replicates previous studies of teaching evaluations. While some of our findings corroborate previous studies, some of our findings contradict existing work. For example, we found no difference in how female TAs were evaluated from how male TAs were evaluated. Previous work on this has found biases against female instructors [16]. As women form a minority of students in CS, we find it even more notable to have found equality in this regard.

That undergraduate TAs were rated more highly in CS1 than graduate TAs is not a new finding by itself [5]; however, that this effect disappears in CS2 is a new finding. We posit that in CS2 and later courses, the greater disciplinary knowledge that graduate students have becomes an advantage for them; in CS1, the institution-specific knowledge that UTAs have, particularly of the structure of how the course works, gives them an advantage, as does their smaller age gap. Future work is needed to try to properly identify why UTAs perform so well in CS1.

The lack of link between experience and evaluations has been previously noted in the literature [16]. The link between student grades and teacher evaluations has been found to be weak [16], and it is not surprising to us that it would disappear altogether for teaching assistants, who are only responsible for a fraction of the student's final mark.

While no prior research has examined the link between instructor workload and their evaluations, prior work has found that instructor workload does affect teaching performance [18]. That this would translate into an effect in evaluations is hence reasonable. What is interesting is where the threshold is for teaching assistants: three lab sections a week appears to be the limit for most TAs, with four sections being clearly too much.

The effect that paired TAs are evaluated as one unit is a new finding. While we cannot explain whether it is due to teamwork, or students not differentiating the TAs, we plan to pursue this in future work. Preliminary work in interviewing paired TAs about their practices does reveal that they work as a team [11].

As for retention, that only the lower cluster of TAs would have an effect on retention is also a new finding. It is unknown whether this is due to low-performing TAs scaring away students from CS, or unengaged students marking their TAs down.

## 6.1 Threats to Validity

While we have a large sample size, our low threshold for significance means that we will not find weak relationships in our data. Our low significance threshold, however, ensures a low likelihood of of false positives.

By using data from only one institution, we can make comparisons within our own context. It allows us to control for factors such as class size, which are consistent between CS labs. But for greater generalizability, replicating our work at other institutions would be the next step.

In our study of TA experience we could not match TA evaluations between courses – only how many times a given TA had taught a given course. This was a result of the way that TA evaluations had been anonymized. Furthermore, our data were restricted to the time periods we had evaluations for – so, our tallies of in-course experience did not include any times a TA may have taught the course before our data begins. Because of this, when correlating the number of previous times a TA had taught the course, we excluded the first two years of our data, since experienced TAs in those years would not have their experience counted.

Our data set is limited to the years in which we have TA evaluations, presenting a threat to validity in calculating TAs' experience. Finally, we only have data for that one set of five criteria (preparedness, etc) – while we can show that these criteria are highly correlated to one another, we do not have comparative data for other possible criteria.

## 6.2 Future Work

As already noted, we plan to further explore the effect of pairing on TAs. Other areas of future work would be examining other quantitative criteria, and qualitative evaluations.

Preliminary data from interviewing TAs about their experiences has found that they do use their teaching evaluations to assess their work and to change their teaching. However, from these interviews, it appears that it is not these quantitative questions, but instead the qualitative portion of the evaluations which gives the TAs the greatest insight about what they should change in their work [11].

## 6.3 Suggestions for Instructors

While these teaching evaluations did provide only a coarse view of how students perceived their TAs, they can still be used by TAs to evaluate their performance. Alone, they do not provide a lot of data – as we note, the qualitative evaluations appear to be what prompt TAs to make change [11]. Yet, at our institution, student evaluations are the only source of formal feedback for our TAs. To complement these evaluations, we suggest to give TAs feedback *from instructors*; preliminary data finds that this is useful for TAs [11]. Peer evaluation may also prove useful for TAs.

In terms of TA management, we recommend avoiding

putting TAs on more than three labs a week. We also advocate favouring UTAs on first-year courses. We advise pairing TAs, especially when worried about weak TAs. Finally, we suggest that for hiring and award decisions, that TAs be judged on more than quantitative evaluations.

# 7. CONCLUSIONS

We found several things in our study:

- The five criteria used to evaluate TAs at our institution all strongly correlate to each other. These evaluations are hence very coarsely grained.
- Undergraduate TAs are ranked more highly overall than GTAs, especially for CS1. The UTAs' popularity, however, decreases with the level of the course.
- There is no difference in how female and male TAs are ranked; students are not biased in this regard.
- We found no correlation between TA experience and their evaluations.
- We found that TAs teaching three or four sections in a term had lower evaluations than those who taught one or two sections.
- Paired TAs were ranked similarly; it appears that either students do not distinguish the two TAs, or, alternatively, the TAs work together as cohesive teams to an extent where TA performance varies by whom they are paired.
- There is no correlation between student final grades and TA evaluations.
- While there is no correlation between highly-ranked TAs' evaluations and how many more computer science courses their students take, there is a negative correlation for poorly-ranked TAs.

The results paint a very coarse picture of TAs – so coarse that it appears that students may not even distinguish the two TAs who teach in a room together. We do, however, have evidence that four lab sections overburdens a TA to an extent where it impacts their evaluations.

We note that a lot of work should be done to improve the usefulness of TA evaluations. More fine-grained detail would provide formative assessment for TAs. Examining alternate criteria would be useful, as would exploring qualitative evaluations. By providing TAs with more useful evaluations, we can assist them to improve their teaching, producing a more effective learning environment for our students.

# 8. ACKNOWLEDGMENTS

For assistance in data acquisition: Colleen Diamond, Chelsey Maher, Joyce Poon and Nasa Rouf. For assistance with the manuscript: Steve Easterbrook.

# 9. REFERENCES

[1] S. S. Bomotti. Teaching assistant attitudes toward college teaching. *Review of Higher Education*, 17(4):371–393, 1994.

[2] M. De Paola. Does teacher quality affect student performance? evidence from an italian university. *Bulletin of Economic Research*, 61(4):353–377, 2009.

[3] J. Felton, J. Mitchell, and M. Stinson. Web-based student evaluations of professors: the relations between perceived quality, easiness and sexiness. *Assessment & Evaluation in Higher Education*, 29(1):91–108, 2004.

[4] S. M. Hobson and D. M. Talbot. Understanding student evaluations: What all faculty should know. *College Teaching*, 49(1):pp. 26–31, 2001.

[5] M. Mendenhall and W. R. Burr. Enlarging the role of the undergraduate teaching assistant. *Teaching of Psychology*, 10(3):184–185, 1983.

[6] V. Muzaka. The niche of graduate teaching assistants (GTAs): perceptions and reflections. *Teaching in Higher Education*, 14(1):1–12, 2009.

[7] D. H. Naftulin, J. John E. Ware, and F. A. Donnelly. The Doctor Fox lecture: A paradigm of educational seduction. *Journal of Medical Education*, 1973.

[8] L. Ni, M. Guzdial, A. E. Tew, and T. McKlin. A regional professional development program for computing teachers: the disciplinary commons for computing educators. *AERA*, 2011.

[9] C. O'Neal, M. Wright, C. Cook, T. Perorazio, and J. Purkiss. The impact of teaching assistants on student retention in the sciences: Lessons for TA training. *Journal of College Science Teaching*, 36(5):24–29, 2007.

[10] C. Park. The graduate teaching assistant (GTA): lessons from north american experience. *Teaching in Higher Education*, 9(3):349–361, 2004.

[11] E. Patitsas. Teaching Assistants in Computer Science: Supporting Their Growth as Teachers, 2012. In preparation.

[12] C. Paul, E. West, D. Webb, B. Weiss, and W. Potter. Important types of instructor-student interactions in reformed classrooms, 2010. American Association of Physics Teachers Summer Meeting.

[13] L. Prieto and E. Altmaier. The relationship of prior training and previous teaching experience to self-efficacy among graduate teaching assistants. *Research in Higher Education*, 35:481–497, 1994. 10.1007/BF02496384.

[14] E. Roberts, J. Lilly, and B. Rollins. Using undergraduates as teaching assistants in introductory programming courses: an update on the stanford experience. *SIGCSE Bull.*, 27(1):48–52, Mar. 1995.

[15] R. Varma. Making computer science minority-friendly. *Commun. ACM*, 49(2):129–134, Feb. 2006.

[16] H. K. Wachtel. Student evaluation of college teaching effectiveness: a brief review. *Assessment & Evaluation in Higher Education*, 23(2):191–212, 1998.

[17] B. C. Wilson and S. Shrock. Contributing to success in an introductory computer science course: a study of twelve factors. SIGCSE '01, pages 184–188, New York, NY, USA, 2001. ACM.

[18] M. Winzer. Grade inflation: An appraisal of the research. Technical report, University of Lethbridge Faculty of Education, 2002.

[19] A. C. Worthington. The impact of student perceptions and characteristics on teaching evaluations: A case study in finance education. *Assessment & Evaluation in Higher Education*, 27(1):49–64, 2002.

[20] S. Wright, A. Wong, and C. Newill. The impact of role models on medical students. *Journal of General Internal Medicine*, 12(1):53–56, 1997.