Unravelling student evaluations of courses and teachers

Antonio Reverter¹, Cristina Martinez², Phil Currey², Severine van Bommel² and Nicholas J. Hudson²,*

Abstract: There is debate over the functional basis of student evaluations of academics, and fresh potential for looking at the data in new ways. Student evaluation data was collated over a three year period (Semester 2 2015 to Semester 1 2018). We used a General Linear Model to estimate the variation in course scores explained by a number of coordinator and course attributes. Three significant factors collectively explain 49% of the School’s variation in course scores—individual coordinator, student evaluation response rate, and mode of delivery. Next, we used hierarchical clustering to explore the inter-relationships among the eight course and teaching evaluation questions. Learning appears to be related to stimulation, whereas overall satisfaction appears to be related to quality of learning materials and course structure (i.e. aspects of course organisation). Student evaluation response rate is positively correlated to all eight course questions, but most positively to a question relating to receiving adequate feedback. This perhaps implies some reciprocity in the flow of information between student and coordinator. The overall teaching rating awarded to academics clusters most with approachability and encouragement of student input—aspects of temperament and style—and not with explanatory skill or organisational ability.

ABOUT THE AUTHOR
Nicholas J. Hudson is a teaching and research academic with broad interests in biochemistry, metabolism and functional genomics. He likes to work across production and wildlife species because this provides broader perspective, embeds observations in an evolutionary context, helps delineate fundamental biological themes and facilitates the disentangling of biochemical signal from noise. The idea for this pedagogical research project was stimulated by his analytical experiences as a genomics scientist. Hierarchical clustering methods are a common tool to help interpret complex multivariate data in the genomics world but have not commonly been applied in the student evaluation context before. Fundamentally, this research was undertaken as the University sector in general debates the usage and implementation of student evaluations. We anticipate the main findings to be relevant to any University that uses similar benchmarking tools.

PUBLIC INTEREST STATEMENT
Retrospective student evaluations of academics and their courses are widely used in Universities as part of teaching and learning quality assurance practices, including performance management. However, there is ongoing debate over their functional basis. We have used some analytical methods not commonly applied before on student evaluation data. This has helped us better understand the basis of our student evaluations within the context of a particular School in an Australian University. We anticipate our findings being applicable to other Schools and Universities who use a similar instrument.

*Corresponding author: Nicholas J. Hudson, School of Agriculture and Food Sciences, The University of Queensland, Gatton Campus, Gatton QLD 4343, Room 224, Building 8117A, Australia
E-mail: n.hudson@uq.edu.au

© 2020 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.
1. Introduction

1.1. Background
Retrospective student evaluations of academics and their courses are widely used in Universities as part of teaching and learning quality assurance practices, including performance management. There is a substantive literature on the pros and cons of these metrics and the extent to which they are considered to provide useful pedagogical information. There have been previous attempts to unravel which factors drive student perceptions (Denson et al., 2010; Ginnse et al., 2007; Hornstein, 2017; Jameson Boex, 2000; Marsh, 1984) however there is some disagreement in this area of research. For example, some authorities consider student evaluations to be (at least partially) uninformative for reasons such as small or non-random student samples (Hornstein, 2017) or because students are not well placed to judge accuracy, depth and currency of academic content (Oermann, 2017). On the other hand, others have argued over an extended time frame that they are a very useful source of pedagogical information with high reliability, stability and validity (Marsh, 1984, 1987; Richardson, 2005).

A study performed on individual level student evaluations within the University of Queensland’s School of Economics found that clear presentation and explanation, in addition to well organised classes were key determinants of instructor effectiveness (Alauddin & Kifle, 2014; Kifle & Alauddin, 2016). According to a review article (Benton & Cashin, 2012) most pedagogical studies conclude that evaluations of teaching are not affected by the teacher’s age, gender, race, personal characteristics and research productivity. Having said this, other studies have found instructor personality (Kim & MacCann, 2018), age (Wilson et al., 2014), race (Smith & Hawkins, 2011) and gender (Boring, 2017; MacNell et al., 2015) to be relevant factors in driving student perceptions. Further, some authorities have found smaller classes (Annan et al., 2013), larger classes (Haslett, 1976) or extreme class sizes (small and large as opposed to medium) (Marsh et al., 1979) tend to award higher evaluation scores. It has also been observed that graduate courses tend to rate higher than lower level undergraduate courses (Annan et al., 2013). Overall, it appears that student evaluation of both courses and individual teachers is complex and there are examples of ambiguous or even conflicting findings.

1.2. Context
With this in mind, we were interested in exploring two facets of student evaluations, in both cases with the overarching objective to better understanding the basis of both the course (and therefore to some extent the coordinator) and teaching evaluation scores within the School of Agriculture and Food Sciences at the University of Queensland. Firstly, we wished to determine the fundamental drivers of a particularly important course evaluation question (Q8 “Overall, how would you rate this course?”) that is used as a proxy at School and Faculty level for whether a given course can be considered overall satisfactory. We used a General Linear Model and examined a range of potential factors including properties of both the course coordinator (gender, rank) and the course itself (class size, campus, delivery mode, number of contributing academics, evaluation response rate) that were of interest.

Secondly, we wished to explore the inter-relationships between the eight evaluation questions of both course and teaching evaluations to develop fresh insight into their functional basis. For this aspect we deployed a hierarchical cluster analysis, widely used in multivariate data analysis in biology, but not commonly applied in the particular context of student evaluations. Hierarchical cluster analysis is a method of cluster analysis whose aim is to objectively find groups within complex data, and whose output is represented as a tree structure called a dendrogram. It provides a formal yet intuitive way of identifying, visualising and interpreting patterns of inter-relationships in multivariate data. We believe our study will contribute to the field by shedding light on the functional basis of student evaluations, particularly for those institutions that use a similar instrument.
2. Materials and Methods

2.1. Human ethics clearance
This study was performed under The University of Queensland’s Human Ethics application “Understanding the basis of variation in student evaluation scores in the Faculty of Science at the University of Queensland” approval 2019001782.

2.2. SECaT question data
At the University of Queensland individual students are asked to retrospectively evaluate courses (Table 1) and teachers (Table 2) based on eight questions. Students assess performance prior to the final grade being awarded so final grade is not a material factor in driving the scores. However, at the point of completing the survey students will have been awarded some proportion of the assessment marks contributing to their final grade. Students respond to each question on a 5 point Likert scale that clearly outlines the levels of increment per scale; selecting from “Strongly agree (5)”, “Agree (4)”, “Neutral (3)”, “Disagree (2)”, and “Strongly disagree (1)”. This ordinal data is reported by the University of Queensland on an interval (means plus standard deviation) basis. Analysis (Cronbach’s alpha) within the University of Queensland suggests that the scale has a high degree of reliability and that there is some redundancy between questions. Reviews discussing this area can be found here (Carifio & Perla, 2008; Norman, 2010).

Analysis is based exclusively on class means provided to the School by the Institute for Teaching and Learning Innovation (ITaLI) at The University of Queensland. Means are reported when a survey received six or more responses as a minimum threshold. We were not provided with individual student evaluation data and these cannot be reverse-engineered.

### Table 1. The eight evaluation questions used to assess course performance

| Question 1 | I had a clear understanding of the aims and goals of the course |
|------------|---------------------------------------------------------------|
| Question 2 | The course was intellectually stimulating                     |
| Question 3 | The course was well structured                                |
| Question 4 | The learning materials assisted me in the course               |
| Question 5 | Assessment requirements were made clear to me                 |
| Question 6 | I received helpful feedback on how I was going in the course   |
| Question 7 | I learned a lot in this course                                |
| Question 8 | Overall, how would you rate this course?                      |

### Table 2. The eight evaluation questions used to assess teaching performance. There was an amended Q7 piloted in Semester 2 2016 (stream a, “I took advantage of opportunities to learn in this course”; stream b, “I made an effort to learn in this course.”)

| Question 1 | The teacher was well organised                                 |
|------------|---------------------------------------------------------------|
| Question 2 | The teacher was good at explaining things                     |
| Question 3 | The teacher was approachable                                  |
| Question 4 | The teacher stimulated my interest in the field of study      |
| Question 5 | The teacher inspired me to learn                               |
| Question 6 | The teacher encouraged student input                          |
| Question 7 | The teacher treated students with respect                     |
| Question 8 | Overall, how would you rate this teacher                      |
2.3. Institutional information
The University of Queensland is an elite Group of 8 Australian University. The School of Agriculture and Food Sciences (SAFS) is the largest School in the Faculty of Science with a broad teaching portfolio including Animal and Plant Agriculture (and Horticulture), Wildlife Biology, Soil Science, Food Science, Social Science and Agribusiness. The School offers 10 undergraduate degree programs, 6 without honours (Bachelor of Agribusiness, Bachelor of Agricultural Science, Bachelor of Equine Science, Bachelor of Science, Bachelor of Veterinary Technology, Bachelor of Wildlife Science) and 4 with honours (Bachelor of Agricultural Science, Bachelor of Equine Science, Bachelor of Science, Bachelor of Wildlife Science). The Bachelor of Science and Bachelor of Science (honours) each have nominated majors.

Entry requirements are different across the programs. Assessments (number, scope, weightings, due dates), assessment criteria and exams (length, supporting material allowed) are variable and are organised on a course by course basis at the discretion of individual course coordinators, although of course coordinators operate under broad University guidelines. SAFS is unique within the University of Queensland in that its teaching is disseminated rather evenly across two geographically and functionally distinct campuses, one located in a suburb of Brisbane City (St. Lucia 27.4698 S 153.0251 E) and one located regionally (Gatton 27.5554 S 152.3372 E). A small number of courses are taught at both campuses, either by repeat delivery or through use of video conferencing technologies.

2.4. Database preparation
Student Evaluation of Course and Teacher (SECaT) data was analysed for the duration Semester 2 2015 to Semester 1 2018. A master database was produced (de-identified and then submitted as Additional File 1) comprising the columns shown in Table 3. The last six columns (“campus” to “gender”) were derived from connecting to other existing databases. Campus (St. Lucia versus Gatton) and delivery mode (internal versus external) were derived from a course catalogue and scheduling database. The number of academics contributing to a given course was determined from the teaching evaluation data as all contributing academics receive a SECaT evaluation. Gender (male versus female) was inferred from personal knowledge of the staff while rank (Professor, Associate Professor and Other) could be identified as it was included in the name column (data true as at Semester 1, 2018).

This database containing the course evaluations and associated data comprised 485 rows representing 72 academic coordinators and 158 courses. The following information was available in the data set: individual coordinator name, coordinator gender, coordinator rank, class size, broad area of study (discipline), number of contributing academics, campus, response rate and mode of delivery.

2.5. Correlation procedure
Based on the 485 records, we used the Procedure Correlation of SAS 9.4 (SAS Institute, Inc.) to explore all the pair-wise correlation coefficient across the eight Course evaluation questions plus three additional variables we considered as quantitative covariates (Agree Q8, class size and student evaluation response rate).

2.6. General Linear Model of Agree Q8
We used the Procedure GLM of SAS 9.4 (SAS Institute, Inc.) to explain variation in Agree Q8 fitting a General Linear Model using least squares methods. After a preliminary analysis to remove non-significant effects, the final model contained the response rate of a linear covariate as well as the class the coordinator with 52 levels, campus with three levels (St. Lucia, Gatton and external), and the interaction of rank by gender.

Broad area of study or discipline was defined by examining the prelude course codes. We explored the following four groups for whom we had adequate data (AGRC for agriculture, ANIM for animal, FOOD for food and a group we designated as OTHER for a range of uncommon prelude codes for which we had only small amounts of data, for example, PLANT, STAT, LAND and CHEM).
Academic rank was defined by looking at the coordinators title correct as at the time of the analysis (November 2018). These ranks were then assessed at three levels (Professor, Associate Professor and Other). The number of academics contributing teaching into the course was binned into three categories, 1, 2 and 3 or greater.

2.7. Hierarchical cluster analysis
Hierarchical cluster analysis is a statistical approach for finding groups in data that are subsequently represented as a dendrogram. This dendrogram provides an intuitive means of visualising how patterns in multivariate data, such as the 8 course and teaching evaluation questions we are scrutinising here, inter-relate. For the course evaluation data analysis, tab delimited text files were created for each semester independently. Each text file contained nine columns (coordinator and course combined into the first column, followed by 1 column for each of the eight evaluation questions) and as many rows as there were courses taught that semester. All evaluation scores were truncated at two decimal places. Each file was imported into PermutMatrix (Caraux & Pinloche, 2005), normalised on rows and then hierarchical clustering was performed on both rows and columns. The resulting dendrogram was exported and manually explored. To get a stronger sense of the overall strength of the associations we also generated a single dendrogram based on evaluation data from all semesters combined into one file.

For hierarchical cluster analysis of the teaching evaluation data, the analytic process described above was repeated. For both courses and teachers the clustering on columns provides information on the inter-relationships between the eight questions, and the clustering on rows provides information on the inter-relationships among the courses/coordinators and the teaching academics, respectively.

Table 3. Summary structure of the master database

| Column Title         | Description                                                                 |
|----------------------|-----------------------------------------------------------------------------|
| Course Code          | The identifier used for the course across years, semesters and delivery mode |
| Class Number ID      | Unique identifier, year, semester and delivery mode specific                |
| Course Coordinator   | Name of the course coordinator including rank                               |
| Mean Q1              | Mean evaluation score for Q1                                               |
| Mean Q2              | Mean evaluation score for Q2                                               |
| Mean Q3              | Mean evaluation score for Q3                                               |
| Mean Q4              | Mean evaluation score for Q4                                               |
| Mean Q5              | Mean evaluation score for Q5                                               |
| Mean Q6              | Mean evaluation score for Q6                                               |
| Mean Q7              | Mean evaluation score for Q7                                               |
| Mean Q8              | Mean evaluation score for Q8                                               |
| % Agree Q8           | % of students who agree with the Q8 value                                 |
| Total Responses      | Number of class who responded                                             |
| Class Size           | Total class size based on data at the time of grade finalisation           |
| Response Rate        | The number of students who responded relative to total class size           |
| Year                 | Year the course was delivered                                             |
| Semester             | Semester the course was delivered                                          |
| Campus               | St. Lucia or Gatton campus                                                 |
| Delivery Mode        | Internal or external (remote) delivery                                     |
| No Academics         | The number of academics contributing to the teaching of the course         |
| Prelude              | A Course identifier indicating broad field of study                        |
| Rank                 | Rank of academic (Professor, Associate Professor or Other)                |
| Gender               | Male or Female coordinator                                                 |
In an effort to determine how sensitive the clustering output is to the input analytic criteria we explored a number of combinations of distance (dissimilarity) and clustering algorithms within Permut matrix using the S1 2018 course data as the test set. All six combinations of three distance (or dissimilarity) metrics (Euclidean, Chebyshev and Manhattan) and two hierarchical clustering (linkage) metrics (McQuitty’s and Complete Linkage) were compared.

3. Results

3.1. Correlation procedure
The correlation output is summarised in Figure 1. The eight course evaluation questions were all highly positively correlated with each other. When examining class size (−0.24, \(P < 0.0001\)) and response rate (0.44, \(P < 0.0001\)) we find that the most extreme correlation to the eight questions is to Q6 (“I received helpful feedback on how I was going in the course”) in both cases, highlighted with a blue rectangle. The second blue rectangle highlights the high correlation (0.94, \(P < 0.0001\)) between Q8 and Agree Q8. We used Agree Q8 for the downstream General Linear Modelling described below.

The mean and standard deviations for the quantitative variables for the 485 records are tabulated in Table 4.

3.2. General Linear Model of Agree Q8
The modelling data is summarised in Table 5. The three significant factors which collectively explained 49% of the variation in Q8 were individual coordinator, student evaluation response rate and mode of delivery. Courses delivered externally receive poorer scores than those delivered internally. The impact of individual coordinator was assessed for the 52/72 coordinators who had at least three evaluation records in the data.

This model explained 49% of the variation in Agree Q8 (with a coefficient of variation of 23.61 and a root mean square deviation of 17.44).

For the categorical variables, the number of records and % representation for each category are as follows: year and semester, 2015 S2 (n = 76, 16.6%), 2016 S1 (n = 82, 17.9%), 2016 S2 (n = 80, 17.5%), 2017 S1 (n = 77, 16.8%), 2017 S2 (n = 68, 14.9%), 2018 S1 (n = 74, 16.2%); campus and delivery mode, external (n = 78, 17.1%), Gatton campus (n = 244, 53.4%), St. Lucia campus (n = 135, 29.5%); number of academics contributing teaching, 1 (n = 163, 35.7%), 2 (n = 114, 24.9%), 3 or more...
3.3. Hierarchical cluster analysis of the eight course evaluation scores

The hierarchical cluster analysis on columns is robust across all Semesters, with only subtle variation in the dendrogram structure (Additional File 2). Figure 2 shows the clustering for Semester 1, 2018—the most recent semester analysed. The shortest branch length (highest relatedness) is awarded to Q2 (The course was intellectually stimulating) and Q7 (I learned a lot in this course). Q8 (Overall, how would you rate this course?) was most tightly clustered to Q3 (The course was well structured) and Q4 (The learning materials assisted me in this course). Q6 (I received helpful feedback in how I was going in the course) was the question most unrelated to the other questions. This basic clustering structure among the evaluation questions is reinforced when data from all semesters is combined and simultaneously clustered (Additional File 3).

In examining the clustering on rows, there are a number of examples where coordinators do self-cluster despite coordinating different courses with different academic content, different modes of delivery and different student cohorts (Additional File 4). This suggests that some academics present surprisingly stable pictures of their coordination.

3.4. Hierarchical cluster analysis of the eight teaching evaluation scores

Like the course evaluation hierarchical cluster analysis on columns, the teaching evaluation cluster analysis on columns is generally robust across all semesters (Additional File 5). In examining the S1 2018, the most recent data, the shortest branch length (highest relatedness) is awarded to Q4 (The teacher stimulated my interest in the field of study) and Q5 (The teacher inspired me to learn).
Q8 (Overall, how would you rate this teacher?) was most related to Q3 (The teacher was approachable) and Q6 (The teacher encouraged student input). Q8 did not form a tight cluster with Q1 (The teacher was organised) or Q2 (The teacher was good at explaining things). Q7 (Did the teacher treat you with respect?) was the question most unrelated to the other questions in S1 2018, although this final aspect applied to only three semesters and can be considered less robust than the other associations (Additional File 5).

Q3 (The teacher was approachable) and Q6 (The teacher encourages student input) cluster with Q8 (Overall, how would you rate this teacher).
The clusters inferred from the S1 2018 course data were largely robust to the particular choice of distance metric (Euclidean versus Chebyshev versus Manhattan) and clustering algorithms (McQuitty’s criteria versus Absolute Linkage). In all combinations, Q6 is the most independent and Q2 is most closely associated to Q7. However, Q8 associates with Q4 using Euclidean distance and Chebyshev distance but with 2 and 7 using Manhattan distance (Additional File 6).

4. Discussion
In this study our main objective was to develop a better understanding of the functional basis of student evaluation scores of courses and teachers. We used a School in the Faculty of Science at the University of Queensland as a case study. We deployed two well characterised numerical approaches to shed light on this question, one of which—hierarchical clustering—has not commonly been applied before in this specific context.

Based on a General Linear Model we found student satisfaction with our courses (using Course Q8 as our proxy) was significantly associated with just three of the many course and coordinator factors for which we had data: these three significant factors which collectively explained 49% of the variation in Q8 were individual coordinator, student evaluation response rate, and mode of delivery. This finding is arguably consistent with the aptitude of the individual coordinator being very important in driving course evaluations. However, we wish to exercise caution in interpretation. We did not formally disentangle the effect of the course from the coordinator by comparing circumstances where a) the same coordinator coordinates different courses, versus b) the same course is coordinated by different coordinators, an avenue for future work.

Moreover, in our School, a coordinator can potentially do all the teaching in a course, or indeed none of it. This presumably complicates the extent to which a student who is evaluating a course (as opposed to evaluating a teacher) has a particular person in mind when making the evaluation. This complexity in interpretation can be contrasted with the teaching evaluations where—by definition—the students do have a particular individual in mind. Having said this, the cluster analysis does provide multiple examples where coordinators strongly self-cluster despite coordinating courses with entirely different content, modes of delivery, assessments, and student cohorts. This latter result tends to imply that some coordinators present very stable external pictures of their coordination style (for reasons unknown) and that the students recognise this.

Low student evaluation response rate is associated with poor course evaluations, and courses delivered externally tend to receive poorer evaluations than those delivered internally. The response rate finding may reflect student (dis)satisfaction governing their preparedness to actually engage in the evaluation process, while mode of delivery may reflect student concerns over levels of feedback and authentic inter-personal contact when studying courses remotely. Some structural factors we expected might influence our course scores such as class size, campus and discipline turned out to be non-significant factors in the model. Interestingly, student evaluation response rate is positively correlated with all eight course SECaT values, but most positively with Question 6 (Did you feel you received adequate feedback?).

We have interpreted this to mean that there is some reciprocity in the way students engage in providing feedback to the University. If students feel they have been given adequate feedback while studying a particular course, then they become more like to engage in the course evaluation process. This is reminiscent of the idea of “closing the loop” in survey research and also suggestive of the presence of a social contract between academic and student. The importance of “closing the feedback loop” in higher education has previously been emphasised (Watson, 2003), but in the slightly different context of making students aware of how their feedback has led to meaningful changes. The argument stated by (Watson, 2003) is that “any improvements that can be made in closing the loop will improve
the likelihood of students providing feedback in the future.” With this in mind, we see enhanced provision of feedback as a fruitful area for improving the student experience in general and for better engaging them in the evaluation process in particular. Class size is negatively correlated with all eight Course SECaT questions, but most negatively with Q6 (Did you feel you received adequate feedback?). This implies that coordinating courses with large class sizes is particularly demanding and that provision of adequate (detailed, personalised?) feedback is problematic in these circumstances. Our results here are broadly in line with observations made previously (Annan et al., 2013).

The hierarchical cluster analysis found very robust (i.e. persistent) inter-relationship among questions for both course and teaching evaluations over the three year period of investigation despite different student cohorts and some staff turnover during this time frame. This implies we have detected meaningful, stable signals despite the generally very high absolute correlations detected among many of the eight questions. With regard to the course evaluation questions, Q8 (Overall, how would you rate this Course?) was most tightly clustered with Q3 (The Course was well structured) and Q4 (The learning materials assisted me in this Course). These are clearly aspects of design and organisation, some of which may take place prior to teaching activities commencing. Interestingly, core educational goals regarding “learning” formed a separate cluster: Q7 (I learned a lot in this course) clusters most tightly with Q2 (The course was intellectually stimulating). Q6 (I received adequate feedback) was the most independent branch in the course tree and it received the lowest mean score of the eight questions.

Analysis of teaching evaluations resonates with our course evaluations in the sense that learning (Q5) and stimulation (Q4) remain very consistently and tightly clustered. Interestingly, the overall teaching rating awarded to academics (Q8) clusters most with approachability (Q3) and encouragement of student input (Q6)—aspects of the teacher’s temperament and style—and not with their explanatory skill (Q2) or level of organisation (Q1). Here, Q7 (The teacher treated me with respect) tended to be the most independent branch in the tree, but this did not apply to all semesters studied. A word of caution: it is pertinent to exercise discretion when interpreting any of the dendrograms as one cannot infer causality. For example, learning and stimulation which are tightly clustered in both course and teaching evaluations could be mutually reinforcing; that is, stimulated students learn more but learning also feels stimulating—rather than being directional.

5. Conclusions
Collectively, we have made inroads into our objective of understanding the basis of student evaluation of course and teaching evaluations in the School of Agriculture and Food Sciences within the Faculty of Science at The University of Queensland. Only a relatively small number of the total factors we explored (individual coordinator, response rate and mode of delivery) possess significant relationships to overall Course satisfaction (Q8). The absence of a gender effect in our course evaluations is in line with the review of (Benton & Cashin, 2012), but in our case the extent to which a coordinator might be considered the “face” of a course varies widely; as some coordinators do all the teaching, some do none, and some do an intermediate amount in co-teaching arrangements.

At this point in time, it is not clear whether there is a gender effect in our teaching evaluation data, as we restricted the General Linear Model analysis to the course evaluations only. Consequently, we view a deeper exploration of our teaching evaluation data as a productive avenue for future work. Another opportunity for developing a deeper understanding of our data would be to explore those cases where an academic receives both coordination and teaching evaluations. In practical terms, one could potentially evaluate academics based on how these two performance criteria diverge and direct them into more or less course management (i.e. coordination) versus face to face teaching. The absence of a class size effect in our course evaluations is at odds with (Annan et al., 2013). The use of hierarchical clustering to untangle the inter-relationships...
among the eight evaluation questions for both courses and teachers is quite novel. This approach highlighted robust inter-relationships across a number of different student cohorts. Our main conclusions here are that features apparently more important for good courses (organisation, quality of learning materials) reflect quite different academic skills to those required of good teachers (approachability and encouragement of student input).

Acknowledgements
We wish to acknowledge Le hou Phan for her advice and comments on an earlier version of this manuscript.

Funding
No funding was received for this study.

Author details
Antonio Reverter¹
E-mail: Tony.Reverter-Gomez@cssiro.au
Cristina Martínez²
E-mail: cristina.martinez@uq.edu.au
Phil Currey³
E-mail: p.currey@uq.edu.au
Séverine van Bommel⁴
E-mail: s.vanbommel@uq.edu.au
Nicholas J. Hudson⁵
E-mail: n.hudson@uq.edu.au
ORCID ID: http://orcid.org/0000-0003-4221-0690
ORCID ID: http://orcid.org/0000-0002-7782-9162
ORCID ID: http://orcid.org/0000-0002-3549-9396

¹ Computational and Systems Biology, Agriculture and Food, Commonwealth Science and Industrial Research Organisation, St. Lucia, Brisbane, Australia.
² School of Agriculture and Food Sciences, The University of Queensland, St. Lucia, Australia.

Declaration of interest
The authors have no competing interests to declare.

Data availability
The data upon which this study is based is available on request in de-identified form.

Supplementary material
Supplemental data for this article can be accessed here.

Citation information
Cite this article as: Unraveling student evaluations of courses and teachers, Antonio Reverter, Cristina Martínez, Phil Currey, Séverine van Bommel & Nicholas J. Hudson, Cogent Education (2020), 7: 1771830.

References
Alauddin, M., & Kifle, T. (2014). Does the student evaluation of teaching instrument really measure instructors teaching effectiveness? An econometric analysis of students' perceptions in economics courses. Economic Analysis and Policy, 44(2), 156–168. https://doi.org/10.1016/j.eap.2014.05.009
Annan, S. L., Tratnack, S., Rubenstein, C., Metzler-Sawin, E., & Hulton, L. (2013). An integrative review of student evaluations of teaching: Implications for evaluation of nursing faculty. Journal of Professional Nursing, 29(5), e10–24. https://doi.org/10.1016/j.profnurs.2013.06.004
Benton, S., & Cashin, W. E. (2012). Student ratings of teaching: A summary of research and literature. The IDEA Center.
Boring, A. (2017). Gender biases in student evaluations of teaching. Journal of Public Economics, 145, 27–41. https://doi.org/10.1016/j.jpubeco.2016.11.006
Caraux, G., & Pinloche, S. (2005). PermutMatrix: A graphical environment to arrange gene expression profiles in optimal linear order. Bioinformatics, 21(7), 1280–1281. https://doi.org/10.1093/bioinformatics/bti141
Carifo, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing Likert scales. Medical Education, 42(12), 1150–1152. https://doi.org/10.1111/j.1365-2923.2008.03172.x
Denson, N., Loveday, T., & Dalton, H. (2010). Student evaluation of courses: What predicts satisfaction? Higher Education Research and Development, 29(4), 339–356. https://doi.org/10.1080/07294360903394466
Gins, P., Prosser, M., & Barrie, S. (2007). Students perceptions of teaching quality in higher education: The perspective of currently enrolled students. Studies in Higher Education, 32(5), 603–615. https://doi.org/10.1080/03075070701573773
Haslett, B. J. (1976). Student knowledgeability, student sex, class size, and class level: Their interactions and influences on student ratings of instruction. Research in Higher Education, 5(1), 39–65. https://doi.org/10.1007/BF00991959
Hornstein, H. A. (2017). Student evaluations of teaching are an adequate assessment tool for evaluating faculty performance. Cogent Education, 4(1), 1304016. https://doi.org/10.1080/2331186X.2017.1304016
Jameson, Boex, L. F. (2000). Attributes of effective economics instructors: An analysis of student evaluations. Research in Economic Education 31(3), 211–227. https://doi.org/10.1020/00220480009596780
Kim, L. E., & MacCann, C. (2018). Instructor personality matters for student evaluations: Evidence from two subject areas at university. British Journal of Educational Psychology, 88(4), 584–605. https://doi.org/10.1111/bjep.12205
MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. Innovative Higher Education, 40(4), 291–303. https://doi.org/10.1007/s10755-014-9313-4
Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. Journal of Educational Psychology, 76(5), 707–754. https://doi.org/10.1037/0022-0663.76.5.707
Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. International Journal of Educational Research, 11(3), 253–388. https://doi.org/10.1016/0883-0355(87)90001-2
Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). Class size, students' evaluations, and instructional effectiveness. American Educational Research Journal, 16(1), 57–69. https://doi.org/10.3102/0021823816010057
Norman, G. (2010). Likert scales, levels of measurement and the “laws” of statistics. Advances in Health
Oermann, M. H. (2017). Student evaluations of teaching: There is more to course evaluations than student ratings. Nurse Educator, 42(2), 55-56. https://doi.org/10.1097/NNE.0000000000000366

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: A review of the literature. Assessment and Evaluation in Higher Education, 30(4), 387-425. https://doi.org/10.1080/02602930500099193

Smith, B. P., & Hawkins. (2011). Examining student evaluations of Black college faculty: Does race matter? Journal of Negro Education, 80(2), 149-162.

Watson, S. (2003). Closing the feedback loop: Ensuring effective action from student feedback. Tertiary Education and Management, 9(2), 145-157. https://doi.org/10.1080/13583883.2003.9967099

Wilson, J. H., Beyer, D., & Monteiro, H. (2014). Professor age affects student ratings: Halo effect for younger teachers. College Teaching, 62(1), 20-24. https://doi.org/10.1080/87567555.2013.825574