2023

A Survey To Evaluate Laboratory Activities Across An Undergraduate Engineering Degree Programme: Data From Five Years Showing Repeatability And Sensitivity

Peter JOHNSON
Imperial College London, United Kingdom, peter.johnson@ic.ac.uk

A survey to evaluate laboratory activities across an undergraduate engineering degree programme: data from five years showing repeatability and sensitivity (PRACTICE)

P. B. Johnson[1]
Imperial College London
London, UK
ORCID 0000-0001-7841-691X

Conference Key Areas: *Innovative Teaching and Learning Methods, Curriculum Development*
Keywords: *Laboratory, evaluation*

ABSTRACT

Laboratory activities are an essential part of an undergraduate engineering education. This paper focuses on evaluating the student experience of laboratory activities. We present a laboratory-specific survey used with large cohorts of students (200) about laboratory activities across an undergraduate Mechanical Engineering degree programme. The key question we try to answer is whether the results of the survey can be used to inform teaching decisions such as which activities need improvement; how to improve them; and to validate these interventions.

We present nine common questions that were used to evaluate activities across a programme. We present five years of data for five of the activites we assess – specifically those from the first year of the programme. The data covers pre-pandemic, lockdown, and post-lockdown periods. The data includes activities that have remained consistent, and activities where changes have been implemented.

For consistent activities, data show good repeatability, adding confidence to the method. The effects of interventions can also be detected. We define a significant change as being a multiple of the standard deviation, across years, when no interventions were used. We discuss the validity of the survey and conclude that, in practical terms, it is useful for informing teaching decisions.

[1] *Corresponding Author (All in Arial, 10 pt, single space)*

P. B. Johnson

peter.johnson@ic.ac.uk

1 INTRODUCTION

Laboratory activities are an essential part of an undergraduate engineering education. This paper addresses the challenge of *evaluating* laboratory activities. The purpose of a high-level evaluation across a degree programme is as follows. Firstly, to provide evidence of continued high quality to justify maintaining the status quo if applicable. Secondly, to identify opportunities for improvement and hence inform the rational allocation of teaching resource to these areas. For example buying new equipment or paying for staff time to improve materials or processes. In the second case, a closer look would be needed before making decisions; the purpose of the high level survey is to identify where attention is needed. Thirdly, evaluation can validate interventions.

Evaluation can consider different perspectives and stakeholders, for example an assessment of learning gains (Watai et al. 2007, Salim et al. 2013), process (Kotulski 2010), or student views (Stark 2016). The student view has been shown to play an important, if not essential, role in evaluating laboratory activities (Nikolic 2016). This paper presents and analyses a survey to evaluate the student experience of laboratory activites.

In this paper we briefly review literature on student evaluations. We present a survey methodology. Example results from the first year of a degree programme are presented, and we discuss the interpretation, validity, and application of the survey.

2 LITERATURE REVIEW

Student evaluation of teaching (SET) is well studied but controversial. A review by Stark (2016) identified the invalidity of surveys – due to bias – of student *judgement* (such as instructor effectiveness or professionalism), but stated that it may be valid to evaluate a student's own experience. We therefore avoid seeking to identify e.g. 'good teachers', but try to identify good (or bad) 'experiences'. A good experience is not the same as effective learning but it is a useful proxy that can be considered along with other types of evaluation.

The rich literature on evaluation tends to focus on classroom activities which are distinct from laboratory activities so not directly applicable. Evaluation of laboratory activities in general is challenging because there is also a diversity in the purpose and nature of laboratory activities across different disciplines and institutional cultures (Feisel and Rosa 2005). Despite this diversity, a common approach to evaluation across a programme is desirable for efficiency in deployment across a programme, and for higher utility – such as rational teaching resource allocation within a programme.

A small number of engineering laboratory-specific surveys are described in the literature. Nikolic (2016; see Ch.3, Ch.7 and Table 7-II) developed a common survey for laboratory activities. The survey used six questions, focussed on 'impression', content/information, 'worthwhile learning experiences', suitability of computers and other equipment, and condition of the laboratory. Repeatability was reasonable, and interventions could be detected. Most questions rely significantly on student judgements.

Corwin et al. (2015) evaluated course-based research experiences in biology, focussing on three dimensions: ‘collaboration’, ‘discovery’, and ‘iteration’. They used a 17 item survey after reducing from a larger bank of questions by testing for statistical validity. The strengths of Corwin et al. (2015) are the statistical validation of the survey; and the identification of key ‘dimensions’, which is a transferable notion and is experience-orientated as opposed to judgement orientated.

In the case of undergraduate engineering education the emphasis is slightly different to the course-based research that Corwin et al. (2015) evaluated so it cannot be directly applied. In this paper we present and analyse a survey that is ‘dimension’ and ‘experience’ orientated like Corwin et al. (2015), but practical in the sense of Nikolic (2016) by having a small number of questions and being applied across a whole programme. The key question we try to answer in this paper is whether the results of such a survey can be used to inform decisions about teaching.

3 METHODOLOGY

We surveyed students using nine standard, closed questions applied to all laboratory activities in their programme. The survey contains supplementary questions that vary year-to-year, and free text comments. In this paper, however, we focus only on the nine standard questions in the survey.

3.1 Nine standard questions

Due to the application of the survey across many activities, we limit ourselves to nine questions to avoid survey fatigue in the students. One student will answer nine questions per activity, for example first year students provide 45 total responses; for our second year students it is 72, which we judge to be a practical limit.

The ‘questions’ are positively worded statements listed in Table 1, answered with a 5-point Likert scale: strongly agree / agree / neutral / disagree / strongly disagree.

Table 1: the nine statements in the 2023 survey.
** = different in previous years; + = clarified from previous years.*

Short name	Full statement
Purpose	‘The purpose of the lab was clear to me.’
Conceptual	‘The lab session gave me a better understanding of the abstract concepts taught in the related module (e.g. energy, pressure, stress, entropy, current, resistance, etc.).’
Challenge	‘The lab session challenged me in a positive way’
Tech. comm. skills+	‘Preparing a report/presentation helped me develop technical communication skills’
Documentation and guidance*	‘The documentation and guidance for the lab session was clear, organised and well prepared. Consider: - Handouts, videos, interactive content - Guidance in the live sessions - Using the practical equipment (where appropriate)’
Engagement	‘I felt engaged in the experience and enjoyed the lab session’
Support*	‘I was well supported by GTAs or other staff during the lab session.’
Feedback	‘I received good feedback [link to College page about feedback] When answering this question consider feedback during lab sessions as well as the marks on your final report/presentation.’
Collaboration*	‘I was able to collaborate with my colleagues (to the extent required).’

Table 2: four statements that changed between years, and their previous versions. Multiple versions indicate multiple changes over different years.

Short name	Previously
Technical comm. skills	**Professional skills** – 'The lab session and report writing helped me develop professional skills' (2019,2020) – 'Preparing a report/presentation helped me develop professional skills' (2021)
Documentation and guidance	**Equipment** – 'I understood how the equipment worked and was able to use it as required.' (2019, 2020)
Support	**Motivation** – 'The lab session was motivating.' (2021) **Practical skills** – 'I learned new and useful practical skills' (2019, 2020)
Collaboration	**Knowledge** – 'I learned new things and reinforced what I had learned in the related module' (2019, 2020)

3.2 Selection of questions

The original nine questions were selected as a starting point based on practical experience, i.e. without a theoretical basis. Annual reviews with a team of activity leaders led to action plans, but we also discussed whether questions should be changed. The changes we made are given in Table 2 and are discussed here briefly.

A survey question about using equipment was refined to focus specifically on documentation and guidance – because this feedback is more actionable for teachers. A question about gaining knowledge overlapped with the question about 'conceptual' understanding, so was replaced by a question focussing on collaboration – a topic that we had previously omitted but was particularly sensitive during lockdown. The question about support reflects a growing focus on the support staff – typically PhD students – and replaced a question on 'motivation' which showed a strong correlation with 'engagement' so was judged to be redundant.

The nine themes that we present here partially correspond to dimensions that have been used in the literature. For example, collaboration (Corwin et al. 2015) and challenge (Kandiko Howson and Matos 2021) are common themes. There is a difference in the overall collection of themes used here because the focus is on discrete activities and on the laboratory. It would require further research to provide a rigorous justification for our 'dimensions'. Currently they have emerged from practice.

3.3 Analysing repeatability and sensitivity

Likert scores, s, range from 1 to 5. For large samples we can consider mean (average) values, μ, of the Likert scale (Derrick and White 2017, 2),

$$\mu = \frac{1}{n}\sum_i^n s_i \qquad (1).$$

where n is the number of responses and i is an index spanning all responsess. We also take an interest in the 'mean of means' μ_2 across years or across the dimensions of an activity. To analyse the repeatability, we consider the standard deviation, σ, of the values of μ over a range of m years:

$$\sigma = \sqrt{\frac{1}{m-1}\sum_j^m (\mu_j - \mu_2)^2}, \qquad (2).$$

where j is the year. More sophisticated approaches, such as omitting neutral responses, and other more advanced filters, did not have a significant effect on the results hence we use the simplest approach – given here – for clarity.

When the difference in the mean, between cohorts or between activities, is greater than the standard deviation ($\Delta\mu \gg \sigma$) we consider the student experience to have changed significantly, as opposed to there being variation between cohorts and/or due to sampling. 'Sensitivity', which is the smallest change, $\Delta\mu$, we can detect above the noise, is therefore an arbitrary multiple of σ, for example 3σ is a practical threshold.

3.4 Context

The survey was optional and anonymous for students and conducted in May each year after all the relevant activities were complete and feedback had been delivered. We conducted the survey from 2019 onwards and present data from 5 consecutive years. Response rates are given in Table 3. For brevity in this paper we only analyse activities from the first year of study. All activities were in-person except in 2021 when they were remote. The survey was conducted at a time when students could socialise, except in 2020 and 2021 when students were in lockdown while filling out the survey. During lockdown (2021) each activity was adapted to remote conditions differently, as summarised in Table 4.

Table 3: response rates and cohort sizes in the survey over five years.

2019	2020	2021	2022	2023
127/174 (73%)	89/172 (52%)	112/214 (50%)	74/199 (37%)	118/194 (61%)

3.5 Description of the activities

All activities involve a scheduled 2-3 hour activity in a laboratory, summarised in Table 4. Some activities involve mandatory preparation work. Most activities involve submitting work a week later to be assessed.

Table 4: Activities in year 1 of study

Activity (subject)	**Purpose and description**	**Lockdown (2021)**
Fairground (solid mechanics)	Introduction to lab work and report writing. Scenario about a fairground ride and drilling a hole. Measure strain on a plate. Write a report.	Data provided
Pipe flow (fluid mechanics, aka 'fluids')	Scenario about a customer complaint. Discovery based learning, testing pipes for pressure drops. Write a report.	'Human robot' in the lab live on Teams following instructions.
Steam plant (thermodynamics, aka 'thermo')	Practical experience of thermodynamic cycles. Measure performance of a steam plant. Consultancy scenatio. Group presentation.	Data provided
Mechatronics	Practical training on a series of DC circuits. Build, test and complete in-lab. Complete before leaving the lab.	Human robot
Materials	Tensile test of aluminium allous; hardness test of carbon steels. Write a report based on a template.	Data provided

4 RESULTS AND DISCUSSION

Fig 1. shows results for five consecutive years, for four of the five different activities in the first year of the programme. The top row of Fig. 1 shows the two higher

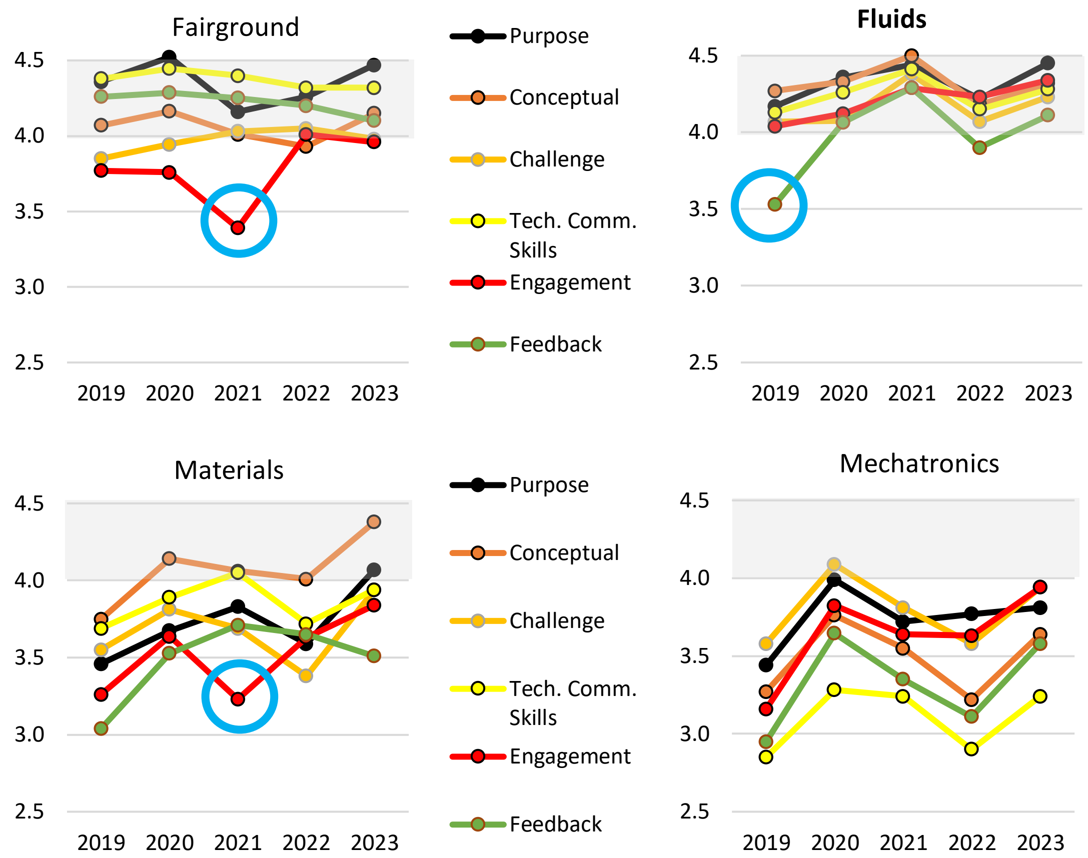

Fig 1. Plots of the mean (μ) in six categories, for four activities, over five years. The legend is the same for all plots. Cyan circles highlight cases for discussion.

scoring activities that were renewed before the survey began. The left column of Fig. 1 shows the activites that provided students with data during lockdown, while the right column shows activities for which the 'human robot' approach was used to acquire data (see Table 4). Further results are provided in Table 5 which includes the standard deviations.

4.1 Repeatability

Consider the 'Fairground' and 'Fluids' activities (top row of Fig. 1, top two rows of Table 5), which both scored consistently high ($\mu > 4$) over the five years. Their standard deviations are $\sigma < 0.15$ in all categories – indicated by green highlights in Table 5 – except two cases, circled in cyan in Fig. 1 and highlighted cyan in Table 5, to be discussed later. The value $\sigma < 0.15$ is significant; for example, a difference $\Delta\mu > 0.5$ between activities or between years is common. Hence variation of the results for these activities is low enough for the survey to be useful in practice.

For the remaining three activities, the standard variation was higher, mostly $0.15 < \sigma < 0.3$. The survey is still useful for these cases, but with lower confidence.

4.2 Sensitivity to difference between activities

The variability for one activity across cohorts (σ), is significantly smaller than the differences in mean scores ($\Delta\mu$) between activities. Perhaps the most obvious result

Table 5: Mean and standard deviation, $\mu \pm \sigma$, for the six consistent questions. Highlights: green: $\sigma < 0.15$, yellow: $\sigma < 0.3$, grey: $\sigma > 0.3$, cyan: notable cases.

	Purpose	Conceptual	Challenge	Tech. Comm. Skills	Engagement	Feedback	μ_2
Fluids	4.3 ± 0.13	4.3 ± 0.12	4.2 ± 0.14	4.3 ± 0.11	4.2 ± 0.12	4.0 ± 0.29	4.2 ± 0.1
Fairg'nd	4.4 ± 0.15	4.1 ± 0.10	4.0 ± 0.08	4.4 ± 0.05	3.8 ± 0.24	4.2 ± 0.07	4.1 ± 0.1
M'tronics	3.8 ± 0.20	3.5 ± 0.24	3.8 ± 0.22	3.1 ± 0.21	3.6 ± 0.30	3.3 ± 0.30	3.8 ± 0.3
Materials	3.7 ± 0.24	4.1 ± 0.23	3.7 ± 0.22	3.9 ± 0.15	3.5 ± 0.27	3.5 ± 0.26	3.6 ± 0.2
Thermo	3.5 ± 0.33	3.8 ± 0.14	3.8 ± 0.18	3.9 ± 0.19	3.2 ± 0.13	3.3 ± 0.10	3.5 ± 0.2

to make this point is for 'Tech comm skills' in yellow in Fig. 1 and the fourth column in Table 5. The mean for the Mechatronics lab, $\mu = 3.1 \pm 0.2$ is at least four standard deviations lower than the lowest scoring of the other activities. This reflects the fact that there is no technical communication in the mechatronics activity – it is a 'build and complete' activity with no assessed report/presentation, while the other four activities have an assessed technical communication. This result validates sensing differences in student experience between activities.

4.3 Validating interventions

The Fluids activity was revised for 2019. The survey indicated that the revision was successful except for feedback, identified in cyan in Fig 1 and Table 5. An intervention was implemented in the following years: feedback sessions between student and marker were scheduled after written feedback had been provided. This practice was copied from the Fairground activity, and the feedback score for the Fluids activity subsequently increased and remained high. The increase of $\Delta\mu = 0.50$ was $> 3\sigma$ in the following four years, which validates the intervention.

A second intervention was in activities that were adapted during lockdown – see Table 4. The left column of Fig. 1 shows activities where data was provided *a priori* to students. Significant drops ($\Delta\mu \approx 0.4$) in engagement (red line) with are evident, and are identified in Fig. 1, left column, by cyan circles. The right column of Fig. 1 shows where a 'human robot' was used to give students agency and interactively gather data. In those cases there is no drop in engagement. The survey provides evidence that a 'human robot' approach was more engaging that providing the data.

The Materials and Thermodynamics labs were subject to many smaller interventions over the years, which may explain the large variations. The Mechatronics lab was essentially the same over five years, so the larger variations for that activity are harder to explain.

5 DISCUSSION

Three aspects of the work presented are discussed here in more detail. Firstly, an interpretation of the results and some clarifications on their meaning. Secondly, a discussion of the validity and limitations of the survey. Finally, some 'use cases' where the results have had an impact on practice.

5.1 Interpreting the results

The relatively low standard deviation ($\sigma < 0.15$) for the two high scoring and consistently delivered activities gives us confidence to attribute larger changes, for example $\Delta\mu > 0.5$, to a change in the student experience. We can in turn, through our wider knowledge of the activities, attribute interventions as causes.

The case of mechatronics illustrates that nuance is required in interpreting the survey. The activity was delivered consistently for five years. The standard deviation was higher ($0.2 < \sigma < 0.3$), but the data in Fig. 1 show that the relative scores of each dimension were consistent. The variation from year to year appears to be uniform (the same for each dimension). In this case, the magnitude of change required in the survey results, before making conclusions with high confidence, is higher. However, the use of relative changes, of a smaller, magnitude can be used in as an alternative metric, albeit with lower confidence.

For any of the survey results, we urge caution when interpreting them. For example the low score in mechatronics for technical communication skills reflects the intent of the teacher. The low score is not necessarily a negative outcome. With this caution in mind, teachers can confidently take action based on the survey results.

5.2 Other forms of validity

Four forms of validity are briefly mentioned here: face, content, criterion, and construct validity. We believe that the face validity of the survey is high because it was developed iteratively through practice and the results are used by practitioners. The content validity is subject to the limited number of questions, for example there are no questions on practical skills. The criterion validity is good in the sense that relative measures are strong, but there is no absolute measure of any of the dimensions. We have not formally considered how students interpret the questions on the survey, but the free text comments and focus groups, combined with our hands-on teaching experience, provide some confidence in this area.

The construct validity of the survey has not been proven statistically. There is evidence in the data that survey results distinguish adequately. Questions were removed when correlations were found. The cases of technical communication skills in Mechatronics, feedback in Fluids, and engagement during lockdown, all support the use of the survey with confidence. This confidence is useful in practice, but it is not a rigorous statistical analysis. The selection of questions is heuristic and would benefit from a more rigorous analysis. Such analysis, however, is likely beyond the scope of this practical survey covering many activities.

5.3 Impact on practice

We use the survey to identify areas where action is needed. We propose an action, identify resources needed to take the action, and indicate *a priori* what changes to the survey results would validate the intervention. We conduct this process as a team across thirteen activities, including the five reviewed in this paper, and another eight in the second year of the programme.

The clearest example of an impact on practice is on the top-right plot in Fig. 1, where a low score for feedback prompted a change to feedback practices and a subsequent, significant, and sustained improvement in the survey scores and follow-up focus groups.

Another example of impact on practice is when an academic member of staff expressed an intention to improve the documentation for an activity. In this case the author advised the staff member not to make the improvement, because the high quality videos they had recently created returned the highest score of any activity in the ‘documentation and guidance’ category – a score which has since been maintained. In this case the impact on practice was to save the opportunity cost of doing work that was unnecessary – despite being well-intentioned.

The final type of example is where staff have targeted an improvement in survey scores with a well intentioned intervention, but that the survey results do not validate the intervention, e.g. the change in score is too small to be significant, or may even be lower. These cases are complex and require deeper analysis than the current survey is capable of supporting. In those cases more detailed evaluation and reflection is required.

6 CONCLUSIONS

A survey containing nine standard questions was used for all laboratory activities across a degree programme. This paper summarised the results from five years of data across five different activities in the first year of an undergraduate mechanical engineering degree programme.

The survey can identify the differences in student experience between concurrent activities and between cohorts. Validated examples include whether technical communication skills improved; the quality of feedback; and the level of engagement. We use the survey in practice to identify areas that need improvement; to validate interventions; and to monitor continued high quality activities.

The strength of the work presented here is its application over five years, and across a programme covering multiple activities. The survey has evolved from practice, which gives it an authenticity and face valitiy, but also limits its construct validity. A statistical validation of the question choice would improve the construct validity.

We will continue to use the survey in coming years, and suggest its wider use across our institution. We would be happy to hear from anyone interested in collaborating on the use of the survey.

Thank you to Idris Mohammed for his assistance with the survey over the last five years.

REFERENCES

Corwin, L. A., Graham, M. J., Dolan, E. L., & O'Flaherty, J. E. (2015). Development and validation of the laboratory course assessment survey: a tool to measure three dimensions of research-course design. *Journal of Educational Psychology*, 107(4), 1006-1022.

Derrick, Ben, and Paul White. (2017) "Comparing two samples from an individual Likert question." *International Journal of Mathematics and Statistics* 18, no. 3: 1-13.

Feisel, L. D., and Rosa, A. J. (2005). "The role of the laboratory in undergraduate engineering education." *Journal of engineering Education* 94, no. 1: 121-130.

Kandiko Howson, C., and Matos, F. (2021). "Student Surveys: Measuring the Relationship between Satisfaction and Engagement" *Education Sciences* 11, no. 6: 297.

Kotulski, T., and Murray, S. (2010) "The national engineering laboratory survey." Labshare Project.

Nikolic, S. (2016). "The Role of Student Evaluations in Improving the Engineering Teaching Laboratory." PhD thesis. University of Wollagong.

Salim, K. R., Ali, R., Hussain, N., and Haron, H. (2013) "An instrument for measuring the learning outcomes of laboratory work." In International Engineering and Technology Education Conference, Ho Chi Minh City, Vietnam.

Stark, P. B. (2016) "Expert Report on Student Evaluations of Teaching (Faculty Course Surveys) Prepared for The Ryerson Faculty Association."

Watai, L. L., Francis, S. A., and Brodersen, A. J. (2007), "A qualitative and systematic assessment methodology for course outcomes from formal laboratory work products in electrical engineering," *2007 37th Annual Frontiers In Education Conference - Global Engineering: Knowledge Without Borders*, Opportunities Without Passports, Milwaukee, WI, USA, 2007, pp. F2C-21-F2C-26.