Anal Bioanal Chem (2017) 409:1719–1724
DOI 10.1007/s00216-016-0109-5

Should students be graded on accuracy and precision? Assessment practices in analytical chemical education

David C. Stone [1]

Published online: 29 November 2016

Introduction

The genesis of this article was a comment from a colleague seeking suggestions for a new experiment, who concluded, "And of course we'll grade on accuracy and precision!" My immediate reaction was, "Why 'of course'?" I will be completely honest and admit that I too grade my students—at least in part—on accuracy and precision. Indeed, most analytical instructors do the same, to various extents; an informal show of hands at a recent conference revealed that only a small minority did not grade students on accuracy and precision at all. Further, most of us were also graded on accuracy and precision as students. It might therefore seem odd to question this practice, which amounts to a "rite of passage" for most introductory analytical courses. Yet we subject analytical methods to considerable scrutiny before accepting them as standard practice; should we not do the same for our methods of assessing student performance?

Accuracy (closeness of results to the true value) and precision (closeness of results to one another) are arguably the most fundamental concepts and essential skills in chemical analysis: without precision in a set of measurements, it is very hard to be confident about their accuracy; similarly, precise measurements that lack accuracy have limited utility. Grading students on their accuracy and precision, however, is *not* something we should do simply out of tradition. As physical scientists, we owe it both to our students and to ourselves to have good reasons for how, when, and to what extent we assign such grades should we decide to do so. This review will provide a framework for making such decisions, drawing on the literature in the field of educational assessment. In its doing so, we will see strong parallels with the principles and practices used in quality assurance, method validation, and proficiency testing.

The larger context

Analytical chemical measurement always occur within a larger context: the information required, the nature of the sample, the resources available, the time and cost of the analysis, and so on [1, 2]. Likewise, assessment occurs within the larger context of the course, the program, and ultimately, the institution within which that course operates [3, 4]. What, then, does this larger context suggest that should inform our discussion of whether to grade students on accuracy and precision?

Most readers will be aware of the pressures driving reform within the higher education sector. Funding challenges, the drive for increased participation rates, and the changing nature of the job market put a strong external focus on student engagement, graduation and employment rates, and "value for money." These act to either reinforce or undermine the adoption of evidence-based best practice and curriculum reform, which themselves are driven internally by the desire to promote excellence in teaching and learning [5, 6]. Both have resulted in an increased focus on clearly stated learning outcomes, goals, and objectives at all levels.

As an example, my own institution's purpose statement with respect to teaching states: "The University will strive to ensure that its graduates are educated in the broadest sense of the term, with the ability to think clearly, judge objectively, and contribute constructively to society" [7].

From this, the overall learning expectations for undergraduate science degrees include breadth and depth of knowledge,

✉ David C. Stone
dstone@chem.utoronto.ca

[1] Department of Chemistry, University of Toronto, M5S 3H6 Toronto, ON, Canada

the integration of skills and knowledge over a degree program, and five core competencies [8]: critical and creative thinking; communication; information literacy; quantitative reasoning; and social and ethical responsibility.

Individual programs are required to indicate how these competencies will be addressed within their disciplinary context. Course curricula must further elaborate on how these competencies will be taught and assessed, as well as the level of attainment expected of students for them to pass. Although this may seem far removed from the question at hand, it is necessary that whatever practice is adopted *must* fit within and support this larger context.

It has long been argued (and equally well contested on grounds of cost and efficacy) that practical work is essential in the physical sciences to provide both discipline-specific skills and enhance conceptual understanding [9–11]. The latter depends, to a large extent, on the style of the laboratory experiments performed [12], and although the former appears self-evident, it too depends heavily on actual implementation. If we argue that a particular skill—especially a transferrable one—is essential to disciplinary practice, then it follows that:

- The development of that skill should be clearly set out as a learning goal.
- The importance of that skill within the discipline should be clearly communicated.
- Students should be provided with the instruction, feedback, and repetition necessary to develop that skill.

The issues facing the instructor are therefore as follows: can gains in conceptual understanding and identified skills be quantified and, if so, how and to what degree of resolution between students? This is important both for students—who require feedback on their progress and assurance that grading practices are fair and reasonable—and for the department, to demonstrate that the value claimed for laboratory courses actually exists. In other words, it is important that a laboratory not only contributes to course, program, and institutional learning goals but that it is seen to do so in quantifiable ways.

The immediate context: a first experiment

The undergraduate analytical chemistry curriculum has been the subject of some debate and review, both in the USA [13–16] and in Europe [17]. This has covered both the content and skills required by practitioners of modern analytical chemistry. Yet the question of *how* skill development is assessed within a laboratory setting has not been well addressed. To illustrate this—and help answer the question posed in this article—consider a representative introductory analytical chemistry course comprising lecture and laboratory components (Fig. 1). This starts with the necessary introduction to measurement, including the definitions of accuracy and precision, and the relevant topics in statistics. The first laboratory, typically, is an exercise in pipetting and weighing framed as the calibration of analytical glassware [18, 19]. Subsequent topics and experiments may include stoichiometric analysis (titrations and precipitations), potentiometry, atomic and molecular spectroscopy, and chromatography.

This first experiment serves two purposes: firstly, it provides a reasonably safe, quick, and simple experiment to help put students at ease; secondly, it introduces the fundamental skills necessary throughout the course—namely, the ability to weigh, dilute, dispense, and calibrate with analytical accuracy and precision. Representative student data, collected over a 3-year period, for the calibration of a grade A 10.00-mL transfer pipette are shown in Fig. 2 as the mean and standard deviation for replicate volumes ($n = 5$) determined by the dispensing and weighing of water. The cumulative mean and standard error of the mean were $9.962_6 \pm 0.004_1$ mL (209 students). We will refer to these data throughout the following discussion. Here, it is sufficient to make two important observations:

1. This is typically the *only* time students are focused *solely* on the practical skills that are fundamental to every other experiment in the course.
2. Students in some institutions receive a grade *solely* for their accuracy and precision and, in extreme cases, may be required to use these calibrated volumes for the remainder of the course.

Methods of assessment in education

In educational terms, assessment can be performed for several purposes [3]. These include *placement*, such as directing a student into a remedial or advanced stream; *formative assessment*, which seeks to reinforce successful learning and highlight areas for development; *diagnostic assessment*, which aims to identify persistent barriers to learning; and *summative assessment*, which evaluates cumulative progress in terms of the intended learning outcomes for the course or unit. These have clear parallels with the principles and practice of valid analytical measurement [1, 2]. Placement, for example, is analogous to the preliminary screening used to identify viable analytical methods and identify potential interferences. Similarly, diagnostic assessment is recognizable in the use of check samples and instrument validation protocols. Formative assessment parallels the processes of both method development and laboratory proficiency testing and training. Finally, summative assessment can be considered equivalent to the generation of a certificate of analysis for a complex sample or certified reference material.

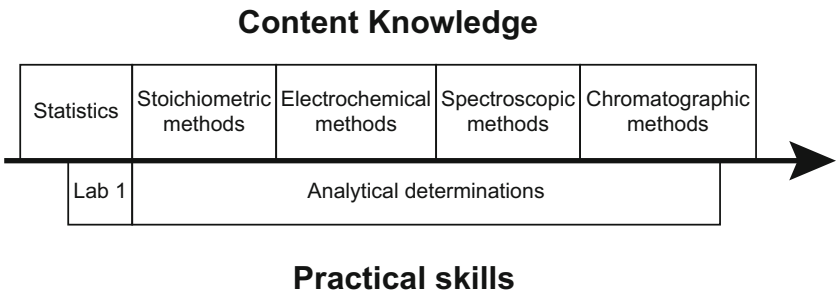

Fig. 1 Structure of a representative introductory course in analytical chemistry

In a laboratory course, grades may typically be assigned for completion of successive experiments, and comprise prelaboratory (quiz), in-laboratory (performance), and postlaboratory (report) components. In practice, such grades often serve as both formative and summative assessment, in the sense that they provide some feedback on each experiment yet *also* count toward the final course grade. How well this works as formative assessment depends on the timeliness and level of feedback provided. If students receive little more than a "B–" a week or more after submitting a report, it is as useful as an analyst reporting a final result as "5.13": without additional context, neither provides any useful information on which to base corrective action. This leads to the first guiding principle for when and how to grade students:

- True formative assessment requires specific and timely feedback, so that students may correct their mistakes *before* the next summative assessment.

For our calibration experiment, feedback may be based purely on each student's mean and standard deviation. But how might we assess student performance in terms of accuracy and precision? In educational terms, evaluations can be either *norm* or *criterion* referenced [3]. As the name implies, norm-referenced evaluations are based on performance relative to the class as a whole, whereas criterion-referenced evaluations are made relative to an independent reference. As an example of the former, we could evaluate precision by assigning numeric grades on the basis of the quartile range for each student's standard deviation (Fig. 3). For the latter, we could similarly assign numeric grades on the basis of deviation from the accepted "true" volume of the pipette (Fig. 4).

It might seem obvious that proficiency should be criterion referenced: successful students should obtain results consistent with the stated volume and tolerance specified by the manufacturer. Yet two issues can undermine this position. Firstly, a student with poor precision may obtain a highly accurate result purely by chance; the 95% confidence interval around that student's mean value belies the agreement of the result with the nominal volume. Secondly, there are multiple sources of error affecting individual results, not all of which may be under student control. For example, improper operation of a mechanical balance by one student can affect scale

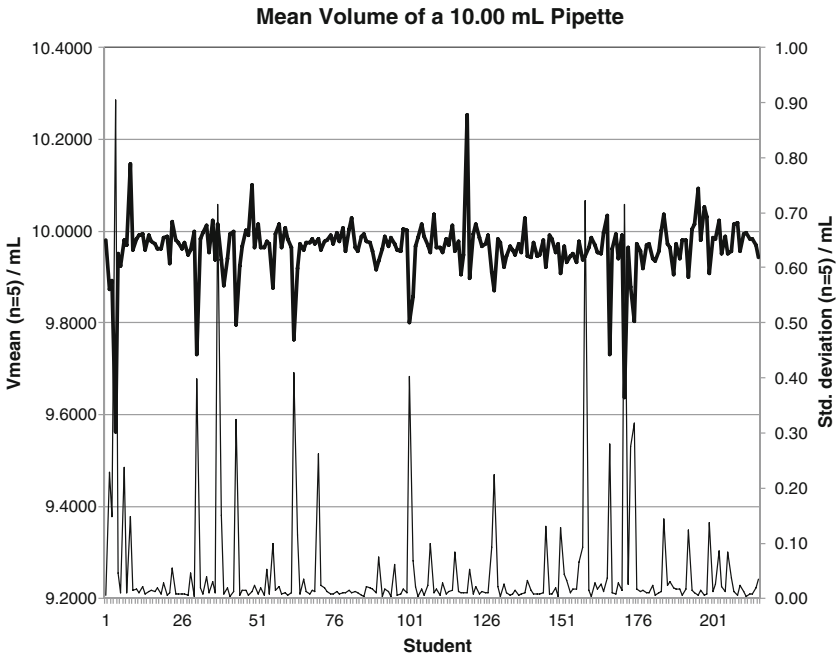

Fig. 2 Cumulative results obtained by students for the calibration of a 10.00-mL transfer pipette by weighing. Primary axis (*thick line*): students' mean reported volume. Secondary axis (*thin line*): students' reported standard deviation for the volume

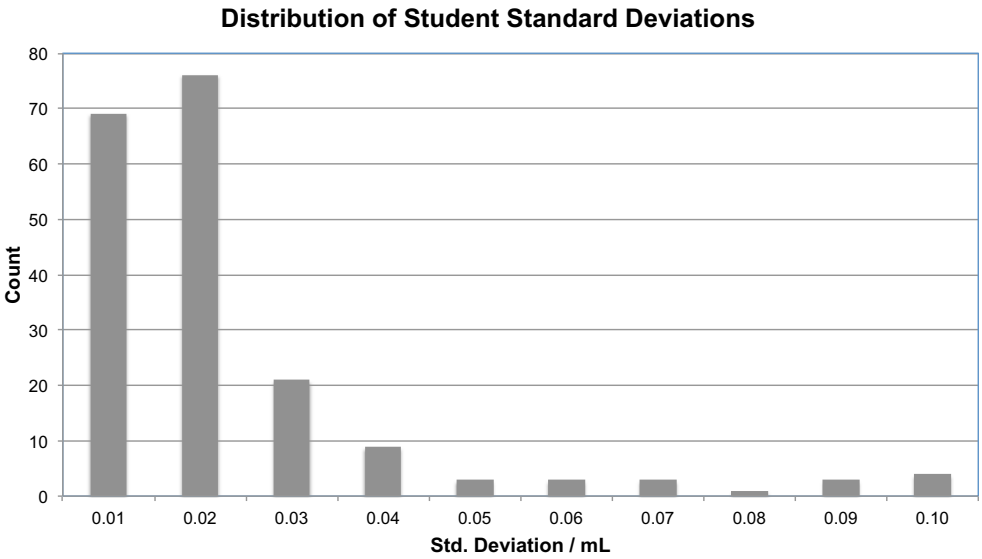

Fig. 3 Distribution of the students' reported standard deviations for the pipette volume

sensitivity, causing errors for all students sharing that balance. Although electronic balances are less susceptible to calibration changes, it is still possible for improper operation to degrade performance for all students.

This second issue can be seen in the class data (Fig. 2), where less than 10% of students reported a mean volume *greater* than the nominal value (10.00 ± 0.02 mL); clearly, the class results show significant bias. For this experiment, students are provided with clean and dry pipettes, record the water temperature, and use the corresponding density in their calculations. Although they do not perform a buoyancy correction (one source of systematic error), this would raise the mean values by only 0.01 mL, much less than the observed bias of 0.04 mL. Although we can explain the remaining bias by common errors in pipetting technique, we could also argue that the laboratory demonstrators should have caught and corrected such errors, and made the students start over.

Validity and reliability in educational assessment

The preceding simple example brings up the issues of assessment *validity* and *reliability* [3], concepts that are highly analogous to their analytical counterparts of accuracy and precision. Validity means that the assessment is both *adequate*—that is, it measures what it claims to—and *appropriate*, meaning that the results of the assessment may be meaningfully interpreted and used as a measure of actual attainment. Reliability refers to the consistency of the assessment, both from one student (or group of students) to another, and on reassessment of the same student(s) on different occasions. In analytical terms, it is a measure of both repeatability and reproducibility. The diagram in Fig. 5 illustrating validity and reliability should look very familiar to analytical chemists everywhere!

Tests, quizzes, and surveys are further evaluated for both *face* and *construct* validity. A multiple-choice question having

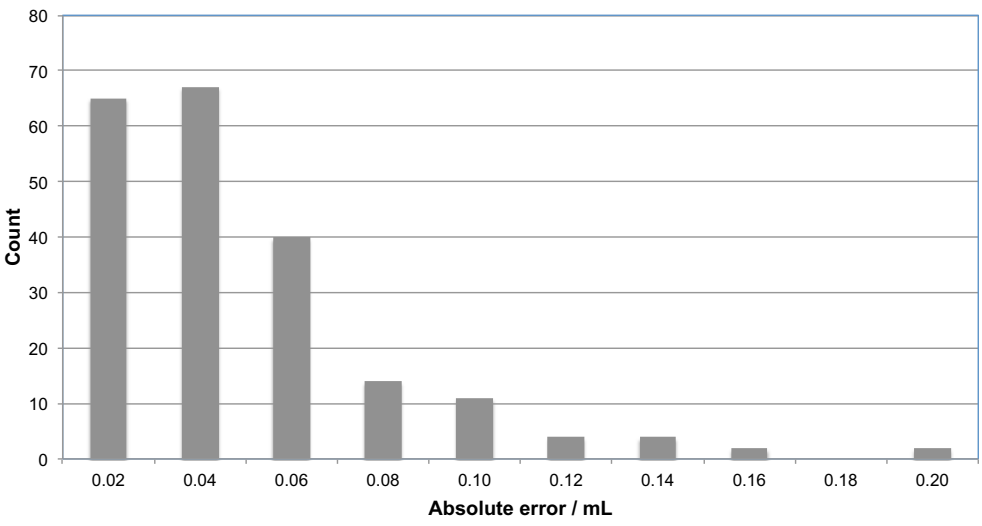

Fig. 4 Distribution of the differences between the nominal volume and student-reported mean volumes

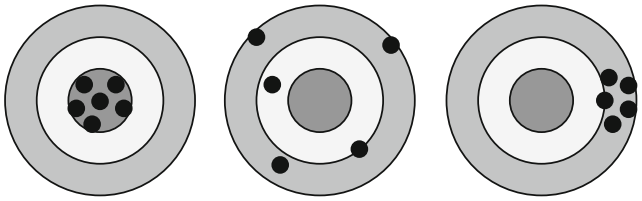

Fig. 5 *Left*: assessments that are both valid and reliable. *Center*: assessments that are unreliable and invalid. *Right*: assessments that are reliable but invalid. (Adapted from [3], Fig. 4.1)

only one correct answer, lacking ambiguity, and marked accordingly would have face validity. If the question were designed to test conceptual understanding but actually tested language proficiency or logical thinking, it would, however, lack construct validity. We can consider this in terms of valid analytical measurement: here, a protocol for measuring iron might have face validity if the results indeed correlated with total iron concentration and not the concentration of other metals, but might lack construct validity if the desired result was the amount of iron(II) and not total iron, or was subject to interference effects that were not accounted for by the method. These considerations lead to a second guiding principle:

- Assessment methods should be *fit for purpose*, demonstrating both *validity* (face and construct) and *reliability.*

Just as we would seek to validate an analytical method by using a standard reference material, comparison with other methods, or interlaboratory trials, so individual assessment exercises should ideally be validated in similar ways. We have already seen how several cohorts of students perform on the pipette calibration exercise relative to the nominal volume and manufacturer's tolerance; however, a more valid performance criterion for evaluation might be the results obtained by the laboratory staff and instructors using the same equipment under the same conditions. Similarly, to be a reliable method of assessment, a student's accuracy should be considered *only* in light of the student's precision.

Reconsidering accuracy and precision

In our deciding the role that accuracy and precision should play within a grading scheme, it is extremely helpful to set out the learning goals and outcomes for the laboratory portion of the course, before we consider how each experiment contributes to meeting those expectations. Some reasonable learning goals would include:

- To understand the importance of accuracy and precision in chemical analysis
- To develop proficiency in the fundamental skills required for chemical analysis
- To correctly apply relevant statistical methods to analytical data to evaluate their accuracy and precision
- To acquire familiarity and practice with common methods of chemical analysis

It should not need stating that setting arbitrary targets for accuracy and precision, then basing student grades solely on these criteria without the opportunity for remediation, does *not* contribute to these goals. If technical proficiency were a requirement, for example, the ideal implementation would involve a formative assessment loop where students would have opportunities to correct mistakes and repeat the experiment until they met or exceeded the required performance standard. This is clearly problematic for large classes with limited time and available resources, which therefore require a different approach. It also leads to a third guiding principle:

- Assessments should be appropriate to the *scale* of the task, including the class size, the time required for adequate assessment, and the available instructional resources.

One approach to accuracy and precision grades for large classes would be a weighting factor that varied throughout the course. Students could then receive feedback without being unfairly penalized for a lack of prior practice. A complementary approach would allow repetition of foundational experiments for students performing below the required standard. In both cases, students need more specific feedback than mere numeric grades. This can be facilitated by careful examination of common student errors and how these influence the final results for each experiment in turn. Suggestions for improvement can then, to some extent, be provided automatically from student data (although insight from direct observation is always more valuable.)

Another approach would be to provide *only* formative feedback for most experiments, and hold periodic laboratory examinations for summative assessment. This is obviously much more time-consuming, but has two potential benefits: firstly, the low-stakes environment can lower student anxiety about the laboratories, resulting in improved learning; secondly, since students have the opportunity to practice essential skills beforehand, they are being assessed more on their acquired proficiency and less on their experience before the course.

Conclusion

Should we grade students on accuracy and precision? Along with the guiding principles outlined earlier, the following points should be considered in answering this question for each specific course:

- Is technical proficiency a reasonable and clearly stated learning goal?

- Are the requirements for formative assessment met by the grading criteria?
- Are policies and procedures in place that allow students to correct poor technique and develop technical proficiency?
- Is each experimental method sufficiently robust that poor accuracy and precision are attributable *only* to student technique?
- Is the method used to assign accuracy and precision grades both valid and reliable?
- Does the weight of grades for accuracy and precision reflect the relative importance of technical proficiency within the overall course learning goals?

Clearly, different conclusions can be drawn for different courses: the requirements of a course for clinical laboratory technicians, for example, should arguably be more stringent than those for a terminal course in introductory analytical chemistry for premed students. Regardless, any laboratory assessment of student accuracy and precision must have demonstrated validity and reliability within the scope and scale of the immediate course context.

The same principles familiar from analytical method development—and the same statistical tests and measures—can be used to examine the validity and reliability of our assessments in making such determinations. Methods of both analytical measurement and educational assessment have a larger objective, must be fit for purpose and appropriate to scale, be robust, and have demonstrated validity and reliability. In short, when it comes to grading students, we should practice what we teach.

Compliance with ethical standards

Conflict of interest The author declares that he has no conflict of interest.

References

1. Tyson JF. Analysis: what analytical chemists do. Cambridge: Royal Society of Chemistry; 1988.
2. Hardcastle WA. Qualitative analysis: a guide to best practice. Cambridge: Royal Society of Chemistry; 1988.
3. Miller MD, Linn RL, Gronlund NE. Measurement and assessment in teaching. 10th ed. Upper Saddle River: Pearson Education; 2009.
4. Diamond RM. Designing and assessing courses and curricula: a practical guide. 3rd ed. San Francisco: Jossey-Bass; 2008.
5. Cooper MM. The case for reform of the undergraduate general chemistry curriculum. J Chem Educ. 2010;87(3):231–2.
6. Cooper MM. Evidence-based reform of teaching and learning. Anal Bioanal Chem. 2014;406:1–4.
7. University of Toronto. Mission. 2016. https://www.utoronto.ca/about-u-of-t/mission. Accessed 9 Aug 2016.
8. University of Toronto. Faculty of Arts and Science. Degree level expectations for honours bachelor degrees. University of Toronto. 2015. http://vpacademic.utoronto.ca/wp-content/uploads/2015/08/dle-fas-hbsc.pdf. Accessed 9 Aug 2016.
9. Hofstein A, Lunetta VN. The role of the laboratory in science teaching: neglected aspects of research. Rev Educ Res. 1982;52(2):201–17.
10. Johnstone AH, Al-Shuaili A. Learning in the laboratory; some thoughts from the literature. Univ Chem Educ. 2001;5:42–51.
11. Hofstein A, Lunetta VN. The laboratory in science education: foundations for the twenty-first century. Sci Educ. 2004;88(1):28–54.
12. Domin DS. A review of laboratory instruction styles. J Chem Educ. 1999;76:543–7.
13. Wenzel TJ. A new approach to the undergraduate chemistry curriculum. Anal Chem. 1995;67:470A–5.
14. Christian GD. Evolution and revolution in quantitative analysis. Anal Chem. 1995;67:532A–8.
15. Perone SP, Pesek J, Englert P. Transforming traditional sophomore quant into a course on modern analytical science. J Chem Educ. 1998;75(11):1444–52.
16. American Chemical Society. ACS guidelines & supplements. 2016. https://www.acs.org/content/acs/en/about/governance/committees/training/acs-guidelines-supplements.html. Accessed 3 Aug 2016.
17. Salzer R. Eurocurriculum II, for analytical chemistry approved by the division of analytical chemistry of FECS. Anal Bioanal Chem. 2004;378:28–32.
18. Fritz JS, Schenk GH. Quantitative analytical chemistry. 4th ed. Boston: Allyn and Bacon; 1979. Experiments 1 and 6.
19. Harris DC. Quantitative chemical analysis. 8th ed. New York: Freeman; 2010. Experiment 1.

David Stone currently holds a continuing appointment as an associate professor, teaching stream, in analytical chemistry at the University of Toronto. He is originally from the United Kingdom, where he obtained both his BSc degree in chemistry and his PhD degree in analytical chemistry (with Julian Tyson) at Loughborough University. He moved to Toronto in 1988 to take up a postdoctoral fellowship with Michael Thompson, coauthoring several articles and a book on acoustic wave chemical sensors. He began teaching general and analytical chemistry in 1993, becoming a full-time member of the teaching faculty in 2003. Since then, his research focus has shifted to chemistry education, with a strong focus on the high school to university transition. He has published and presented numerous articles and papers on both chemistry education and teaching in analytical chemistry, culminating in the 2015 Chemical Institute of Canada Award for Chemical Education.