frontiers in Education

OPINION
published: 29 January 2020
doi: 10.3389/feduc.2020.00002

Definitions of Formative Assessment Need to Make a Distinction Between a Psychometric Understanding of Assessment and "Evaluative Judgment"

*Anders Jönsson**

Faculty of Education, Kristianstad University, Kristianstad, Sweden

Keywords: formative assessment, evaluative judgment, assessment, psychometrics, Assessment for Learning (AfL)

OPEN ACCESS

Edited by:
Matthew Courtney,
The University of Melbourne, Australia

Reviewed by:
Fei Chen,
University of North Carolina at Chapel Hill, United States
David Leaper,
Macquarie University, Australia

***Correspondence:**
Anders Jönsson
anders.jonsson@hkr.se

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement,
a section of the journal Frontiers in Education

Received: *18 July 2019*
Accepted: *14 January 2020*
Published: *29 January 2020*

Citation:
Jönsson A (2020) Definitions of Formative Assessment Need to Make a Distinction Between a Psychometric Understanding of Assessment and "Evaluative Judgment". Front. Educ. 5:2. doi: 10.3389/feduc.2020.00002

INTRODUCTION

Definitions of formative assessment have probably always been too broad (for an overview, see Cizek et al., 2019, p. 7). When trying to encompass every conceivable way of using assessment data to support student learning, some ways that are not very likely to have any positive effects will also be accommodated by the definition. Although this may not be a reason for invalidating a definition, it may have a problematic influence on pedagogical practice. In Sweden, for example, the practice of adjusting teaching based on last year's results on the national tests has been called formative assessment, although the students that actually did the tests are not affected by the use of assessment data. Another common example is when students fail a teacher-made test and are allowed to take another similar test, until they–either by chance or acquired knowledge—are able to perform above the cut-score.

Given the broad definition of formative assessment, it is not surprising that new definitions emerge, such as the "next-gen definition" in the recently published *Handbook of formative assessment in the disciplines* (Cizek et al., 2019). These authors, however, place the definition of formative assessment within a psychometric understanding of assessment, which may, among other things, exclude formative assessment practices that do not adhere to psychometric ideals. Most tellingly, in a recent opinion paper, Brown (2019) questions whether "Assessment for Learning" is really assessment. His argument is that only those evaluative processes that meet psychometric standards should be termed assessment, which means that the in-the-moment and on-the-fly decisions that are part of "Assessment for Learning," should preferably be called something else.

Although it at first sight may seem controversial to make a distinction between formative assessments and "real assessments," this is not necessarily an unreasonable idea. In 1989, Sadler suggested that teachers' "qualitative judgments" should form the foundation for formative assessment. Qualitative judgment—or "evaluative judgment" as it is now called by those guarding the legacy of Sadler (Boud et al., 2018)—is an evaluative process, which is not only essentially different from the psychometric understanding of assessment, but also more attuned to the in-the-moment judgments made by teachers as an integral part of teaching.

The argument made here is therefore that the next generation of definitions of formative assessment should not only incorporate both evaluative judgments and a psychometric understanding of assessment, but also take into consideration the differences between these evaluative processes, since they can be optimized to support different kinds of decisions.

A PSYCHOMETRIC UNDERSTANDING OF ASSESSMENT

In the *Handbook of formative assessment in the disciplines*, the editors define formative assessment as an "inferential activity":

> In all assessment contexts, conclusions about student learning are based on incomplete and indirect samples of information from which necessarily tentative conclusions are made. /.../ The possible sources of evidence about student learning are many and varied; these sources of evidence must be synthesized to arrive at the judgments that will inform decisions about progress, understanding, next steps, pedagogical choices, and so on (Cizek et al., 2019, p. 14–15).

The view expressed in this quote is a typical example of a psychometric understanding of assessment, which is manifested in the idea that the goal of assessment is to draw conclusions about something that is not visible to the naked eye (i.e., student learning). Since student learning is not visible, these conclusions have to rely on "indirect samples of information." The idea of using indirect measures is commonplace, not only in psychometrics, which may hide the fact that this is not necessarily the only possible route for assessments.

For those who are not familiar with the concept of indirect measures, a short explanation will be provided and one of my personal favorites is the so-called "cloud chamber." The cloud chamber is basically a sealed, but transparent, container filled with water vapor. Since water vapor is not visible to the naked eye, the container looks empty. However, if a radioactive substance emitting ionizing particles is placed within the chamber, mist-like trails of tiny droplets form in the vapor. The cloud chamber thereby acts as a simple particle detector, making the invisible particles observable as trails of water droplets. What is important to remember, however, is that is not the actual particles that can be observed in the chamber, but water. The ionizing particles are hence only observed indirectly, but by the use of appropriate theory, the existence of (non-visible) ionizing particles may be *inferred* from the (visible) water droplets.

In a similar manner, psychological (and invisible) constructs, such as intelligence and understanding, are measured indirectly, mainly through tests. Here, the answers to the test items correspond to the trails of mist in the cloud chamber, and they can be used to make inferences about students' learning or understanding. In order to do this, theory is needed, so that the test scores can be translated into conclusions about student characteristics. Test scores do not speak for themselves, no more than trails of water droplets in a sealed container.

Within this paradigm, each item on a test is used as an indication of a latent (invisible) trait, and more items generally means that more accurate conclusions can be drawn from the test scores. This view is represented in the quote above by Cizek et al. (2019), which states, first, that there are many and varied possible sources of evidence about student learning, and second, that these sources of evidence must be synthesized in order to make sense. A similar view is presented by Brown (2019), who mentions portfolios, authentic assessments, and peer assessment as examples of different methods of "data elicitation" that can be used to make inferences about student learning.

QUALITATIVE JUDGMENT

In the article "Formative assessment and the design of instructional systems," Sadler (1989) defines the concept of qualitative judgment. Qualitative, or evaluative, judgment is used to appreciate *the quality* of student performance. According to Sadler, a qualitative judgment is made by a knowledgeable person and is not reducible to a formula that can be applied by non-experts. Qualitative judgments are made through the use of criteria and typically multiple criteria are used simultaneously when appraising the quality of performances.

There are several important distinctions that can be made between the two perspectives "assessment-as-judgment" and "assessment-as-testing," but the most fundamental is probably that the "focus of assessment" differs. In psychometrics, assessment is an inferential process, since the focus of assessment is an invisible, theoretical construct (such as knowledge or competency). Psychometric assessments therefore rely on indirect measures, using aggregated data from several items. In qualitative judgments, on the other hand, the focus of assessment is quality of performance. By judging the quality of an essay, a report, or other kind of extended task, the assessment is *direct* (Frederiksen and Collins, 1989), which means that *no inferences* have to be made about the student's knowledge or other latent traits (Note the wording "have to be," since inferences *can* be, and often *are,* made in practice. An extreme position is deliberately taken here, however, in order to more clearly distinguish between the two perspectives). Compare, for instance, with the widely used metaphor by Robert Stake, where summative assessment is compared to a guest tasting the soup. In such a case, it is the quality of the soup that is being evaluated, not any characteristics of the cook. Similarly, when an anonymous manuscript is being assessed by a reviewer, it is the quality of the text that is being assessed, and no inferences *need to be made* about the author of the manuscript. In order to make valid inferences about the cook or the author, more than one performance is typically needed.

Another important feature of assessment-as-judgment is that the tasks used to assess student performance are not arbitrary items, which can be traded for other items with similar psychometric properties. Rather, these tasks need to give the students the possibility to show whether they are capable of producing the qualities sought for. For example, in relation to the soup metaphor, the cook has to make a real bowl of soup, otherwise no assessment of the quality can be made. Although an analytic assessment can be made by evaluating individual aspects of the soup (such as temperature, thickness, saltiness, and so on), these aspects have to be evaluated in relation to the whole. Qualities such as thickness and saltiness clearly cannot be evaluated in isolation from the soup! Furthermore, this identification of strengths in relation to certain aspects or criteria, and weaknesses in relation to others, can be used as raw material for formative assessment. Since the assessment is direct, no translation is needed of the assessment outcome (i.e., strengths and weaknesses in relation to criteria) in order to provide constructive feedback to the students. This clearly differs from assessment-as-testing, where the scores from a test have to be translated in order to be used as meaningful feedback. Furthermore, if students are trained in using the same criteria as the teacher, they can develop their own sense of quality, which can be used to self-assess and self-regulate their learning.

Even if evaluative judgment is a compelling alternative for assessing performance tasks, there are limitations to this approach. Most notably, although not the primary focus of this paper, the use of evaluative judgment for summative purposes have been questioned due to sometimes inconsistent and biased judgments. In fact, as noted by Gauthier et al. (2016), rater variability may at times explain more of the total variability than students' own performances. The drive toward competency-based education, relying heavily on rater judgements, has therefore spawned an increased interest in research about "rater cognition" and the process of assessment.

The problems concerning inconsistent and biased judgments are not only relevant for assessments with a summative purpose, however, but can also affect decisions made for formative purposes, such as decisions about assigning harder or easier curriculum materials or identifying specific learning difficulties (Brown, 2019). The defining character for these decisions is that they are based on accumulated evidence about student learning, not the quality of performance on an individual task. This is clearly seen in research on teachers' grading, where intuitive and holistic approaches are used to aggregate multiple sources of evidence into a single grade. Under such circumstances, the assessment is heavily influenced by the idiosyncratic beliefs of individual teachers, resulting in a situation where grades from different teachers differ substantially (e.g., Brookhart et al., 2016; Malouff and Thorsteinsson, 2016). Most interestingly, for the argument made here, is that even in cases where teachers agree on (a) which criteria to use, (b) the strengths and weaknesses in students' performances, and (c) the rank order of students–they still assign different grades (Jönsson and Balan, 2018). This would suggest that the teachers agree on the quality of student performance (i.e., there is consistency in their evaluative judgments), but that they employ different strategies for aggregating this information into a single grade.

DISCUSSION

As can be seen from the presentations of evaluative judgment and assessment-as-testing above, the former is a more compelling foundation for those evaluative processes where teachers make judgments about the quality of student work and provide formative feedback. The main reason for evaluative judgments being more compelling in such situations is that these judgments rely on direct assessments, where the assessment outcome (i.e., strengths and weaknesses in relation to criteria) is directly available to students without the need for translation. This may in turn facilitate the development of students' own evaluative judgment, supported by practice in peer- and self-assessment. Assessment-as-testing, on the other hand, relies on indirect measurements where translations of outcomes are always necessary. The difficulties in involving the students in interpreting the assessment data from such assessments, means that assessment-as-testing is typically teacher centered, rather than learner oriented.

However, that evaluative judgment is a more compelling foundation for those in-the-minute and on-the-fly judgments about the quality of student work does not mean that it is a better foundation for *all* evaluative processes that currently fall under the umbrella term formative assessment. On the contrary, the case has been made that assessment-as-testing is likely to be a more appropriate approach for those formative assessments, where teachers need to make decisions based on accumulated evidence about student learning. Consequently, a standardized and psychometrically sound procedure could support the teachers in reaching consistency in their grading.

What is proposed is therefore that the next generation of definitions of formative assessment include both evaluative judgment and a psychometric understanding of assessment. Furthermore, the distinction between these evaluative processes should be taken into consideration, so that they can be optimized to support different kinds of decisions:

- Evaluative judgments should preferably be used for the day-to-day interactions around the quality of students' performances. This means that evaluative judgments should be seen as a legitimate form of professional practice, and that performance tasks (such as argumentative texts, lab work, and oral presentations) should not be replaced by tests or other standardized assessments based on indirect measures, since this is likely to have a negative impact on students' development of own evaluative judgment and self-regulation of learning.
- Psychometrically sound assessment procedures should preferably be used for those formative assessments where teachers need to make decisions based on accumulated evidence about student learning, so that these decisions do not rely on the idiosyncratic beliefs of individual teachers.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This publication was part of a project supported by the Swedish Research Council (2018–03389).

REFERENCES

Boud, D., Ajjawi, R., Dawson, P., and Tai, J. (2018). *Developing Evaluative Judgement in Higher Education*. London: Routledge.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., et al. (2016). A century of grading research: meaning and value in the most common educational measure. *Rev. Educ. Res.* 86, 803–848. doi: 10.3102/0034654316672069

Brown, G. T. L. (2019). Is assessment for learning really assessment? *Front. Educ.* 4:64. doi: 10.3389/feduc.2019.00064

Cizek, G. J., Andrade, H. L., and Bennett, R. E. (2019). "Formative assessment: history, definition, and progress," in *Handbook of Formative Assessment in the Disciplines,* eds H. L. Andrade, R. E. Bennett, and G. J. Cizek (New York, NY: Routledge), 3–19.

Frederiksen, J. R., and Collins, A. (1989). A systems approach to educational testing. *Educ. Res.* 18, 27–32.

Gauthier, G., St-Onge, C., and Tavares, W. (2016). Rater cognition: review and integration of research findings. *Med. Educ.* 50, 511–522. doi: 10.1111/medu.12973

Jönsson, A., and Balan, A. (2018). Analytic or holistic: a study of agreement between different grading models. *Pract. Assess.* 23.

Malouff, J. M., and Thorsteinsson, E. B. (2016). Bias in grading: a meta-analysis of experimental research findings. *Aust. J. Educ.* 60, 245–256. doi: 10.1177/0004944116664618

Sadler, R. D. (1989). Formative assessment and the design of instructional systems. *Instr. Sci.* 18, 119–144.

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.