ALTERIDAD
REVISTA DE EDUCACIÓN
p-ISSN:1390-325X / e-ISSN:1390-8642
http://alteridad.ups.edu.ec

july-december 2020
Vol. 15, No. 2,263-275
https://doi.org/10.17163/alt.v15n2.2020.10

Validity and reliability in student learning evaluation throughout active methodologies

Validez y confiabilidad en la evaluación del aprendizaje mediante las metodologías activas

Dra. María del R. Medina-Díaz is professor and researcher at Universidad de Puerto Rico (Puerto Rico) (maria.medina2@upr.edu) (https://orcid.org/0000-0002-0197-2480)

Dra. Ada L. Verdejo-Carrión is professor and researcher at Universidad de Puerto Rico (Puerto Rico) (ada.verdejo@upr.edu) (https://orcid.org/0000-0001-9522-3116)

Received: 2020-02-01 / **Revised:** 2020-06-01 / **Accepted:** 2020-06-09 / **Published:** 2020-07-01

Abstract

The demands on university education call for changes on teaching strategies and in the evaluation of student learning. Active methodologies are part of these strategies, which facilitate the development of student learning or competences, through situations or problems close to the real world and to a professional career. These require to rethink, plan and guide teaching as student-centered, as well as to use techniques and techniques for collecting valid and reliable information that leads to an appropriate and a fair evaluation of student learning. However, the evidence of validity and reliability of the interpretations of the scores or information collected with these tools has not had enough attention, according to the literature reviewed. The purpose of this paper is to discuss the validity of the interpretation and reliability of the scores or the information collected through classroom assessment tools in universities. Accordingly, to some publications, a set of recommendations is provided with sources of evidence that underpin validity and reliability. At a minimum, it is suggested taking into account evidence related to content validity and to internal consistency of the scores or the information collected, when making judgments and decisions that affect the students. It is concluded that greater prudence is needed in the interpretations and inferences of learning, if there is insufficient validity evidence.

Keywords: Active methodologies, student evaluation, validity, reliability, student learning, evaluation techniques.

Resumen

Las exigencias en la educación universitaria demandan cambios en las estrategias de enseñanza y en las técnicas y los instrumentos que contribuyen a evaluar el aprendizaje estudiantil. Las metodologías activas, como parte de estas estrategias, facilitan el desarrollo de determinados aprendizajes o competencias, mediante situaciones o problemas vinculados con el mundo laboral y social. Esto requiere replantear, planificar y orientar la enseñanza centrada en el estudiantado y utilizar técnicas e instrumentos para recoger información que conduzcan a emitir juicios apropiados, certeros y justos de los aprendizajes. Sin embargo, no se ha prestado mucha atención a la validez y la confiabilidad de las interpretaciones de las puntuaciones o la información recopilada con estos instrumentos para el uso propuesto, según se desprende de las publicaciones revisadas. El propósito de este trabajo es aportar a la discusión acerca de la validez y confiabilidad de las puntuaciones o la información recopilada con los instrumentos aplicados en las aulas universitarias. Se consultaron varias publicaciones especializadas y se presentan algunas recomendaciones acerca de las fuentes de evidencia para sustentar la validez y la confiabilidad. Como mínimo, se sugiere la evidencia relacionada con el contenido y la consistencia de las puntuaciones o la información, al emitir juicios y tomar decisiones que afectan al estudiantado. Se concluye que se necesita mayor prudencia en las interpretaciones e inferencias de los aprendizajes, si no existe suficiente evidencia de la validez.

Suggested citation: Medina-Díaz, M. del R., & Verdejo-Carrión, A. (2020). Validity and reliability in student learning evaluation throughout active methodologies. *Alteridad, 15*(2), 263-275. https://doi.org/10.17163/alt.v15n2.2020.10

Descriptores: Metodologías activas, evaluación de estudiantes, validez, confiabilidad, evaluación del aprendizaje, técnicas de evaluación.

1. Introduction

Over the past three decades, the demands of the public, the governments and agencies of Higher Education require the student to have varied and complex learning for the performance at work and throughout life (Erwin, 1991; Huba & Freed, 2000; Krzykowski & Kinser, 2014; McClarty & Gaertner, 2015; Pozo & Pérez-Echeverría, 2009; UNESCO, 1998). These expectations involve certain changes in teaching strategies, techniques and instruments that collect information to evaluate the student learning. It involves conceiving learning differently to address the characteristics, processes and styles of learning and provide instructional activities, where students construct knowledge and skills based on the previous knowlege, as well as opportunities to become actively involved, demonstrating what they have learned and evaluating performance (Brookhart, 2004; Erwin, 1991; Hortigüela-Alcalá et al., 2015; Huba & Freed, 2000; López-Pastor & Sicilia-Camacho, 2016). With this perspective, the emphasis is on how to learn or develop mental structures and processes of thinking and acting in order to develop and achieve the expected learning, which are many and integrated into the cognitive, affective, psychomotor and social dimensions of academic and personal development of the student in different educational contexts. Commonly, these learning objectives are set out as teaching targets (Stiggins, 2017, p.11); learning outcomes, "learning goals", (Huba & Freed, 2000, p. 94, p. 9, respectively) and competences (Baartman et al., 2006; De la Orden, 2011; Epstein, 2007; Fernández March, 2006; García-Merino et al., 2016; Goñi Zabala, 2005; Olmos-Miguelañez & Rodríguez-Conde, 2010; Pozo & Pérez- Echeverría, 2009; Voorhees, 2001).

Active or authentic methodologies, such as teaching strategies, act as a vehicle to facilitate development and achievement through situations or problems similar to those faced in professional fields and society. The application of these methodologies requires rethinking, planning and guiding teaching in different ways, aligning techniques and tools for evaluation and considering the student as the focus of the process. A technique to evaluate refers to the set of procedures or actions planned to collect information about learning, while an instrument or tool is the specific object or medium to apply it.

The information collected covers scores, selections, annotations, comments or other ways that require responses or observations. Medina-Díaz and Verdejo-Carrión (2019) classify them into four groups with the associated instruments: (a) tests (e.g., objective and subjective evidence); (b) observation (e.g., checklists, category scales and headings); (c) personal communication (e.g., interview, notebook) and (d) performance tasks (e.g., project, portfolio). Angelo and Cross (1993), Barkley and Major (2016), Suskie (2009) and Weimer (2013) present multiple examples of these assessment techniques for the university context. Professors are expected to know and use those that harmonize with the learning objectives, the teaching strategies employed and the student characteristics, in order to select appropriate information (Banta et al., 1996; Black & William, 1998a, 1998b; Bennett, 2011; Davies & Taras, 2018; Lopéz-Pastor & Sicilia-Camacho, 2016; Newble & Cannon, 1991; Olmos-Miguelañez & Rodríguez-Conde, 2010; Rawlusyk, 2018; Webber, 2012). It should be clarified that a technique and an instrument is not exclusive to a type of evaluation (e.g., diagnostic, formative or summative), but rather it is the professor who determines the purpose and use of the scores or the results. This is, precisely, the validity of the interpretations of the scores or information for the designated use, with the expectation that will substantially improve the quality of the academic experience and learning or competences of the university student. Table 1 presents three examples of active methodologies

and possible information collection techniques and instruments for evaluating student learning (Medina-Díaz & Verdejo-Carrión, 2019).

Table 1. Examples of active methodologies, techniques and instruments to collect information

Active methodologies	Definition	Techniques (and instruments)
Project	A set of activities carried out by the student, individually or in groups, for a long time, for the purpose of dealing with a problem and producing an object, prototype, oral or written report.	Systematic Observation (Checklist) Personal communication Learning log) Performance Task (Rubric)
Problem solution	The process by which the student performs a series of actions and makes decisions integrating knowledge, skills and attitudes to respond to or solve a problem or a real situation and for which there is no single solution.	Personal communication (Reflection and self-assessment forms) Performance task (Rubric)
Cooperative learning	Small group of students in which everyone interacts and participates to help each other understand an issue, perform a task or achieve a common goal.	Systematic observation (Rating scale) Personal communication (Learning log, self-assessment and co-assessment forms)

Student evaluation involves making an informed judgment, based on appropriate and relevant information about various learnings developed and achieved. Therefore, one of its great challenges is to collect and combine information, both quantitative and qualitative, obtained with multiple instruments and at different times. However, under the informality and how quickly classroom assessment often occurs, the validity of interpretations of the scores or information collected, as well as reliability, is not considered. Perhaps for this reason, there is a gap on the research and publications focused on Higher Education Assessment techniques (Angelo & Cross, 1993; Barkey & Major, 2016; Huba & Freed, 2000; Suskie, 2009; Wolf et al., 2012), as well as research on university faculty evaluation practices in several countries (Alquraan, 2012; Andreu-Andrés & Labrador-Piquer, 2011; Bearman et al., 2017; Brown & Atkins, 1988; Erwin, 1991; Gilles, Detroz & Blais, 2010; Goubeaud, 2010; Goubeaud & Yan, 2004; Hernández, 2012; Hortigüela Alcalá et al., 2017; Pereira & Flores, 2016; Yükselii & Gündüz, 2017). It is also possible that, for the sake of trust and academic freedom, validity and reliability are assumed when professors develop or use one or more instruments to apply them to the student, without having sufficient evidence to support them and caring about certain technical elements (Esteve Zarazaga, 2007; Gil Flores, 2005; Jacobs & Chase, 1992; O'Hagan, 2014; Poskanzer, 2002). The evaluation of learning integrated with active methodologies in teaching invites to consider the quality of the instruments applied at the university level, especially the validity and reliability of the scores or the information collected. The main purpose of this work is to contribute to the discussion of these two aspects.

2. Methodology

To this end, several publications that promote or discuss the validity and reliability of interpretations of information collected with the tools used to evaluate learning were reviewed

(American Educational Research Association et al., 2018; Brookhart, 2003, 2007; Cizek, 2009, 2015; Joint Committee on Standards for Educational Evaluation, 2018; Moss, 2003). The Standards for Educational and Psychological Testing, published by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2018) point to the requirements that instruments developed and administered for a large scale or commercial profit must support the validity of interpretation of scores for the proposed uses. They present a framework of reference or guidance to ensure that relevant issues are addressed in the construction of educational and psychological tools and provide a basis for reviewing and criticizing them (American Educational Research Association et al., 2018, p.1). The Joint Committee on Standards for Educational Evaluation (2015), is in charge of this topic "Q4 Reliability and Validity": "Classroom assessment practices should provide consistent, dependable and appropriate information that supports interpretations and decisions about each student's knowledge and skills" (Location 643).

From our perspective, the same statistical breadth and rigor is not projected to occur as in the validation procedures used in standardized utilization tests and other instruments applied in educational and psychological practice and research. However, those used in universities should generate appropriate quantitative and qualitative information about the student learning according to the purpose, content, teaching strategies and size of student groups, as well as inferences and actions derived. One of these inferences could be that the student has achieved problem-solving competency, applying descriptive statistics. Moss (2003) argues that the notion of validity should be reconceptualized for classroom practice, and she advocates interpretive approaches (e.g., based on sociocultural theory and hermeneutics) to handle the information that is collected continuously. Brookhart (2003) proposes the development of a classroom-based measurement theory ("classroometric", p.8). Faced with these approaches, the concepts of validity and reliability are highlighted as well as the sources of evidence needed. In addition, several recommendations are done on the relevant evidence for interpreting scores and other information collected with the instruments.

3. Validity and reliability

3.1. Validity

Validity is defined to the "degree to which evidence and theory support interpretations of test scores for the proposed uses of tests" (American Educational Research Association et al., 2018, p.11). In other words, it implies a judgment about the interpretation of scores or information obtained with an instrument, in the light of evidence from these sources of evidence are based on content, response, processes, internal structure, relation with other variables and consequences (American Educational Research Association et al., 2018). This vision is based on a unifying concept of validity, noting that the integrated evidence from these sources contributes to the validity related to the construct, so that it theoretically and empirically supports that the instrument measures or represents it appropriately and leads to appropriate inferences and actions (Messick, 1989). An instrument used in university classrooms requires relevant evidence to support the interpretations and uses of the scores or information obtained. Table 2 summarizes some of the procedures linked to the five sources of evidence, according to Cizek (2009), Medina- Díaz and Verdejo-Carrión (2019), McMillan (2008) and Nitko and Brookhart (2011).

Table 2. Recommendations on the validity and reliability of the interpretation of scores or information collected with an instrument

Source	Recommendation
Validity related with the content	• Prepare instrument specifications with the content topics. • Provide a sufficient number of items or tasks associated with the content. • Create the instrument, following the recommendations of recognized reference sources in the field of learning assessment. • Review the comprehension of the instrument items or tasks, as well as the instructions. • Ensure that the vocabulary, grammatical structure, language and format of items or tasks are suitable for the student. • Determine the alignment between items or tasks in the instrument with the learning and content.
Validity related to the answer process	• Check the match between the answers offered to items or tasks and learnings (e.g., cognitive strategies or processes). • Identify cognitive processes, skills or strategies needed to answer items or tasks. • Interview a group of students, immediately after answered the instrument, to know the strategies or processes used to respond. • Provide time to respond to or apply cognitive processes or complex skills. • Ask the student to explain the work done or to show the steps or procedures for reaching an answer.
Validity related to the internal structure	• Analyze the consistency of responses to items or tasks of an instrument associated with the same learning target or content. • Check the match between the scores, and the performance previously qualified with an instrument.
Validity related to other variables	• Compare the results of an instrument, before and after the discussion of a topic or teaching-learning process. • Contrast the performance in different instruments that represent the same and different learning. • Identify the characteristics, experiences and educational needs of the student who responds or with whom the instrument is used.
Consequences	• Identify the effects or impact of the use of the instrument and the information obtained. • Review the interpretations and decisions made about learning and instruction, according to the information collected. • Associate the interpretation of the instrument information with the corresponding decisions. • Request the student reactions or comments about instruments and scores.
Reliability	
Consistency in the answers or execution	• Have enough information about the learning developed and achieved through various instruments. • Have a key or guide to review answers to items or the performance tasks. • Establish and report the criteria and indicators that will be used to observe or score performance on tasks. • Provide examples of expected responses and works at different levels of performance. • Review or score all student group answers to one question, before moving on to another. • Provide two or more occasions to answer questions or tasks, related to expected learning and compare the performance. • Use two or more people (professor and student) to observe the performance, compare scores, and calculate the percentage of agreement. • Describe procedures for correcting scoring responses or performance.

The evidence based of content-related validity starts with the identification of the purpose and learning targets intended to be represented in an instrument. The instrument´s construction requires delimiting the learning, content topics, quantity and type of items as part of the specifications that serve as the basis. For example, when it comes to an objective test, a table of specifications relates the learning targets or content topics with the items (Medina-Díaz & Verdejo-Carrión, 2019). In this and other instruments designed such as performance tests, a concordance is expected between items or tasks, objectives, teaching strategies and the emphasis and time spent on content discussion (i.e., instructional validity). The items or tasks are created from the specifications, taking into account the expected learning, the time require to answer, the students´ needs and the university context. To develop essay or discussion-like questions which are very often used in this context, the organization and relationship between topics is also considered in the ideal answer (Brown, 2010; Medina-Díaz & Verdejo-Carrión, 2019).

The items or tasks, along with the instructions, are the main pieces of an instrument. Therefore, their selection and development requires time. Nitko and Brookhart (2011), Haladyna (1997), Mateo and Martínez (2008); Medina-Díaz and Verdejo-Carrión (2019) present different recommendations for creating them. The most general are: (a) clarity in vocabulary used, (b) simplicity in grammatical structure and (c) avoidance of including two ideas, as well as sexist, offensive or discriminatory language. In addition, the correct or expected answers to the items or tasks are written. A crucial issue is to ensure that items or tasks require showing cognitive learning, at least understanding and applying concepts, actions, and procedures. This assumes that professors have used teaching strategies that fostered their development in class.

Also, content organization, edition and physical display are taken into account on the construction of an instrument. This is reflected in the appearance and organization, as well as the absence of spelling errors in the items and instructions. This applies to both a printed and electronic instrument. These actions contribute to the evidence related to the content of the instrument, which is more important in the evaluation of student learning.

It should be remembered that the review of the representativeness and relevance of the items or tasks of a test or other standardized instrument depend on the judgment of people who know or are familiar with the content, and student group of interest (American Educational Research Association et al., 2018). In university classrooms, this work is done only by the professor, who decides about the learning to be represented, the content and the items or tasks of the instrument. In addition, it is possible to under - or over - represent some objective or topic of the course. To minimize this, a colleague could help review the agreement or match of the instrument with the specifications and the ambiguity in the questions, before applying in it.

Evidence based on response processes concerns the fit between instrument items or tasks and the cognitive, affective, and psychomotor learning or skills required to answer or complete them. To gather evidence about different cognitive processes or thinking skills, it is necessary to at least identify how items or tasks model them and get relevant information. For example, a performance task in Mathematics expects that the student uses a general or heuristic strategy (e.g., make a diagram or drawing) to answer it. The professor may interview several students and ask them to explain their reasoning to verify whether they responded by applying this strategy to the task or solution to the problem. The professor can also observe the performance of a group of students carrying out a task to verify the procedure employed.

The evidence based on the internal structure involves the cohesion of the items or tasks of the instrument in representing what is proposed (e.g., learning or content topics), and the con-

sistency in the responses. Nitko and Brookhart (2011) suggest paying attention to patterns of response to items or tasks, as well to the concordance with those managed instruments. On tests, it is advisable to analyze and compare the answers to various questions or tasks of the same content in order to determine the consistency of the results (McMillan, 2008). The evidence of the relationship with other variables is related to the association between the responses to the items and the instrument scores with external variables (called criteria). This evidence is relevant when scores on an instrument serve as an indicator of performance in other variables (e.g., academic average, scores on a logical reasoning test). This could be used in the evaluation of learning, first by identifying the characteristics, experiences and diversity of educational needs of the student who responds or with whom the instrument is used, and then, in seeking similarities and differences in the execution of the instrument (McMillan, 2008). Observations on the student's answers and works could also be compared before and at the end of the discussion. The differences suggest possible changes in learning. It would also be useful to verify the performance on several tasks aimed at demonstrating similar or associated learning.

The evidence related to the consequences considers the possible effect or impact of the instrument and the use of the information collected. It requires documenting how scores were interpreted (e.g., normative- or criterion- referenced), what they were used for and what the consequences were (e.g., increasing motivation or time to study, reducing the number of failures). For example, the impact on the student learning of a performance task aimed at conducting research or practical work in a community could be investigated (Ricoy & Fernández-Rodríguez, 2013). An interview with the student or a written reflection would serve as evidence. McMillan (1997) and Taylor and Nolen (2005) emphasize the consequences for teachers and students; particularly on the effect of feedback, techniques and the instruments administered in the motivation and study habits.

Bonnen (2013), Gipps (1994), Medina Gual (2013), McMillan (1997) and Muñiz and Fonseca-Pedrero (2008) also include a number of proposals to demonstrate the validity of interpretations of the information collected, as well as the reliability. Suskie (2006, p. 37) presents four characteristics of useful instruments (that they produce accurate information; have a clear purpose; engage teachers and students; focus on clear and important learning goals). Medina Gual (2013) proposes a scheme with curricular, interpretative and instrumental evidence and their respective strategies. Gibbs (1994, p. 174) presents six quality criteria of an instrument: (a) curricular fidelity, (b) comparability, (c) dependence, (d) public reliability, (e) description of the context and (f) equity. The equity of fair treatment of students is crucial by proving various opportunities and whatever is necessary to achieve the expected learning. For its part, McMillan (1997, p.49) raises the following: (a) clear and appropriate learning objectives, (b) appropriate assessment methods, (c) validity, (d) reliability, (e) justice, (f) positive consequences, and (g) practicality and efficiency. These latter criteria include several aspects that must be taken into account by professors, such as the time and resources to prepare the instruments, apply it and score the performance of the student, as well as the complexity to interpret the results. For example, creating an objective test takes a long time, but it takes a short time to answer and correct it. A performance task takes more time to create, answer and score its responses For this reason, it is recommended that some items or tasks be reviewed and reused from time to time if its possible.

3.2. Reliability

Reliability refers to the consistency of scores or information obtained with an instrument applied in different times or moments. It is related to the precision of scores or other information

from a group of students with the least possible errors. The errors could be linked to changes in the conditions of administration of the instrument, the subjectivity in the correction or qualification of the scorers, ambiguity in the items, as well as the lack of motivation and the doubt of the student.

The number of items, tasks, and replications when instruments are used is a reliability-related factor. Typically, increasing the number of items in an objective test increases the reliability coefficient. There are three main types of reliability coefficients obtained through statistical procedures: (a) stability (or test-retest), which refers to the consistency of scores over time or at different moments; (b) equivalence, which concerns whether two or more parallel forms of an instrument produce scores or similar results; and (c) internal consistency, which focuses on the cohesion of responses to items on an instrument, which attempts to measure or represent the same objective or content. Theoretically, if an instrument produces reliable scores, these should be similar for the group of students who answer it on two or more times. The correlation between the scores is the stability coefficient.

In most classrooms it is impossible to perform double administration, so reliability based on internal consistency, which requires only one, is used. An objective test has a key for correcting the responses to the items, so the subjectivity in the correction is not a limitation. The professor who applies it to large groups of students and has appropriate computer software to analyze the responses to the items and calculate a coefficient of reliability (usually internal consistency), if the scores are compared with a norm group (norm-referenced interpretation). Also, the standard measurement error can be estimated if desired. This indicates the accuracy in the individual scores of the instrument and it depends on the magnitude of the reliability coefficient and the variability of the scores; i.e., at a higher reliability the lower the measurement error. Often, these statistics are not usually calculated on tests applied in university courses or departments. However, taking them into consideration for reviewing technical quality requires this effort, especially when a test is known as "reliable".

With regard to the performance tasks (e.g., project, portfolio) essentials in the active methodologies, the subjectivity in the scoring is reduced but not extinguished, with the use of a rubric, checklist or rating scale containing the appropriate criteria and indicators (Medina-Díaz & Verdejo-Carrión, 2019; Reddy & Andrade, 2010; Selke, 2013; Van der Schaaf, Baartman & Prins, 2012). In addition, the student must know these in advance or can participate in their elaboration. If possible, it is advisable to provide examples of expected responses, actions or jobs at the different performance levels included in the rubrics. In the absence of these, procedures for correcting or scoring responses are described. If open answer or essay questions are included, then the answers of the entire group of students to one question are scored before reviewing the answers from another. This not only helps maintain consistency in grade, but provide feedback to the student when discussing answers to questions.

In this way, reliability is manifested by consistency in assigning scores with a rubric to score each student's performance in the performance task. This involves two procedures for finding consistency: intra and inter-judge. The intra-judge consistency depends on how the professor applies an instrument (e.g., a rubric), in a stable way to rate the answers or the student's work. For this, the professor can re-evaluate a sample of previously reviewed topics and find the match in the scores, as well as identify the differences (Cizek, 2009). The inter-judge agreement requires two or more people to review and rate execution. This is unusual in classrooms unless there is the collaboration of another professor or student, thus, this is a good opportunity to encourage the participation of the student as observers, or judges as a co-assessment. The professor and one or more students, independently, rate the performance of a student by using the

same instrument. The scores are then compared and an agreement percentage or other statistic is calculated (Stemler, 2004). Disagreements in scores that suggest little or no reliability in scores may be caused by applying the instrument differently or by possible bias in the qualifying person (e.g., lenience or severity). This is another moment to involve the student in the dialogue and experience of qualifying and evaluating, as well as to understand the nature of the process and the common and different performance scores. In addition, professors also benefit from confirming the correct scores obtained.

Moreover, Brookhart (2003) and Smith (2003) define reliability as a sufficiency or abundance of information. Brookhart (2003, p.11) refers to the stability of information to detect the difference between expected and the current state of the student performance or the amount of information. Smith (2003, p.31) refers to obtaining sufficient information, having a complete view of the student, and leading to a good decision. It is also necessary to consider whether the student had several opportunities (items or tasks) and moments to show his/her learning and what he/she is able to do. This allows to observe the variation and consistency in the performance, and thus to formulate better interpretations of the learning achieved. A professor could use this information and a complementary information (e.g., interview) to derive inferences about the learning achieved by the student. It also helps reduce the anxiety or fear it could cause if there is only one time or instrument to demonstrate what has been learned. As can be seen, in these cases no reliability coefficient is calculated. Reliability depends on the use of various techniques that yield consistent information from expected learning. It should be emphasized that having reliable scores or information is not enough to support validity.

Finally, the convenience of using various tools to collect information allows to overcome the limitations each has and try to represent the complexity and multiple learning dimensions. The combination of quantitative and qualitative information accumulated throughout the teaching-learning process offers a more comprehensive and accurate look at the student´s learning and thus allows to make decisions and formulate appropriate and fair judgments. Central tendency and variability statistics are useful for describing the performance of a group if considering scores of a test or rubric applied to a task. These are also shared with the student. Where there is qualitative information (e.g., essay, reflection) or graphic (e.g. comic book, infogram), certain criteria are considered (e.g., vocabulary, argumentation, use of examples and symbols) to describe the performance or development of each student. Analysis strategies can be used to identify common or divergent elements or patterns in written or graphical parts. Information collected with various instruments is combined and compared ("triangulate data") to identify topics or patterns which converge on decisions and judgments about the student´s learning and the instructional process: What learning did they obtain? What difficulties do they present? What is inferred about learning? How will the results be used? What changes are needed in teaching strategies?

4. Discussion and conclusions

Student evaluation is a systematic process of making a judgment based on the information gathered about the learning developed and achieved, throughout the teaching-learning process. The trust placed in the information collected depends on the quality of the instruments that professors create, manage and use. The validity lies in the appropriate and credibleness of the interpretations of the scores or the information collected by an instrument about a student's learning. The evidence gathered through the different sources strengthens the certainty of interpretations and inferences, both of the learning process and of the performance of the student. In addition, it drives the decision-making in the teaching-learning process (e.g., expand

discussion of a topic, offer practical experience, use an active method in teaching or recommend tutoring) and the judgments issued (e.g., "Mary has achieved the competences in Math"). These decisions and judgments also deserve pondering.

As observed, validity and reliability have not had enough attention in the discussion of information-gathering techniques and instruments applied by professors to evaluate student learning. Possibly, they are on a "list of endangered species" as Popham stated (2005, p.71). We hope that this paper will contribute to the reflection and action to preserve them. We suggest more caution in the interpretations of the information acquired with the instruments that are developed and used along with active teaching methodologies if there is no evidence of the validity that supports them. Finally, we recommend that evidence related to the content and consistency of scores or information obtained be taken into account when making judgments and decisions that affect students. However, we recognize the challenge and effor of this proposal in the face of the reality of teaching job and the respect for academic freedom of professors in the different universities. Dealing with the possible fragility of interpretations and decisions derived from the information obtained or accumulated about student learning, more evidence is necessary to support them.

There is no doubt that it is not enough for professors to create and manage better instruments if they do not use the information appropriately, consistently and fairly in the evaluation of student learning. As Palomba and Banta (1999) and Banta and Pike (2012) indicate, the results need to be used to close the assessment cycle (plan-collect information-interpreting it-using the results). Also, information concerning cognitive learnings is not sufficient if affective, psychomotor, social and other that are relevant in the different university disciplines (e.g., safety in the management of materials or substances, autonomy and teamwork) are not taken into consideration. In addition, the evaluation of learning, in an ethical and constructive way, points to the right of the student to be notified about and participate in the process of applying the criteria, techniques and instruments; as well as the interpretation of the information collected, both to improve learning and teaching. It should be remembered that interpretations, inferences and decisions made have consequences (some more serious than others), for the students and the society. Finally, the expectation of improving learning does not depend exclusively on the evaluation, but also in the changes in the vision of learning, teaching strategies, curriculum of disciplines, professional development and leadership of the faculty and the collaboration of the administration of universities to support, integrate and evaluate them.

References

Alquraan, M.F. (2012). Methods of assessing students' learning in higher education An analysis of Jordanian college and grading system. *Education, Business and Society: Contemporary Middle Eastern Issues*, *5*(2), 124-133. https://doi.org/10.1108/17537981211251160

Andreu-Andrés, M.A., & Labrador-Piquer, M.J. (2011). Formación del profesorado en metodologías y evaluación. Análisis cualitativo. *Revista de Investigación en Educación*, *9*(2), 236-245. https://bit.ly/2UhFxu0

Angelo, T.A., & Cross, K.P. (1993). *Classroom assessment techniques* (2da ed.). Jossey-Bass.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (2018). *Estándares para pruebas educativas y psicológicas* (M. Lieve, Trans.). American Educational Research Association (Original work published 2014).

Baartman, L.K.J., Bastianens, T.J., Kirschner, P.A., & Van der Vieuten, C.P.M. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programs. *Studies in Educational Evaluation*, *32* (2), 153-170.

https://doi.org/10.1016/j.studuc.2006.04.006

Banta, T.W., Lund, J.P., Black, K.E., & Oblander, F.W. (1996). *Assessment in practice: Putting principles to work on college campuses*. Jossey-Bass.

Banta, T.W., & Pike, B.R. (2012). The bottom line: Will faculty use assessment findings? In C. Secolsky & B. Denison (Eds.), *Handbook on measurement, assessment and evaluation in higher education* (pp. 47-56). Routledge.

Barkley, E.F., & Major, C.H. (2016). *Learning assessment techniques: A Handbook for college faculty.* Jossey-Bass.

Black, P., & William, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7-74. https://doi.org/10.1080/0969595980050102.

Black, P., & William, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*, 139-144.

Bearman, M., Dawson, P., Bennett, S., Hall, M., Molloy, E., Boud, D. & Joughin, G. (2017). How university teachers design assessments: A cross disciplinary study. *Higher Education*, *74* (1), 49-64. https://doi.org/10.1007/s10734-016-0027-7

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*, 5-25. https://doi.org/10.1080/0969594X.2010.513678.

Bonnen, S.M. (2013). Validity in classroom assessment: Purposes, properties, and principles. In J.H. McMillan (Ed.), *Sage Handbook of Research on Classroom Assessment* (pp. 87-106). Sage.

Brookhart, S.M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice*, *22*(4), 5-12. https://doi.org/10.1111/j.1745-3992.2003.tb00139.x

Brookhart, S.M. (2004). Assessment theory for college classrooms. *New Directions for Teaching and Learning*, *100*, 5-14. https://doi.org/10.1002/tl.165

Brookhart, S.M. (2007). Expanding views about formative classroom assessment: A review of the literature. In J.H. McMillan (Ed.), *Formative classroom assessment: Theory into practice* (pp. 43-62). Teacher College Press.

BrOwn, G.T.L. (2010). The validity of examination essays in higher education: Issues and responses *Higher Education Quarterly*, *64*(3), 276-291. https://doi.org/10.1111/j.1468-2273.2010.00460.x

BrOwn, G., & Atkins, M. (1988). *Effective teaching in higher education.* Routledge.

Cizek, G.J. (2009). Reliability and validity of information about student achievement: Comparing large-scale and classroom testing contexts. *Theory into Practice*, *48*(1), 63-71. https://doi.org/10.1080/00405840802577627

Cizek, G.J. (2015). Validating test score and detecting test score use: Different aims, different methods. *Assessment in Education: Principle, Policy & Practice*, *23*(2), 212-225. https://doi.org/10.1080/0969594X.2015.1063479

Davies, M.S., & Taras, M. (2018). Coherence and disparity in assessment literacies among higher education staff. *London Review of Education*, *16*(3), 474-490. https://doi.org/10.18546/LRE.16.3.09

De la Orden Hoz, A. (2011). Reflexiones en torno a las competencias como objeto de evaluación en el ámbito educativo. *Revista Electrónica de Investigación Educativa*, *13*(2), 1-21. https://bit.ly/2QIILUX

Epstein, R.M. (2007). Assessment in medical education. *New England Journal of Medicine*, *356* (4), 387-396. https://doi.org/10.1056/NEJMra054784

Erwin, T.D. (1991). *Assessing student learning and development*. Jossey-Bass.

Esteve Zarazaga, J.M. (2007). Un examen a la cultura escolar. *Avances en Supervisión Educativa*, (7). https://bit.ly/3ajVMfN

Fernández March, A. (2006). Metodologías activas para la formación de competencias. *Educatio Siglo XXI*, *24*, 36-56. https://bit.ly/39deN20

García-Merino, J. D., Urionabarrenetxea, S., & Bañales-Mallo, A. (2016). Cambios en metodologías docente y de evaluación: ¿Mejoran el rendimiento del alumnado universitario? *Revista Electrónica de Investigación Educativa*, *18*(3), 1-18. https://bit.ly/2WII8yE

Gil Flores, J. (2005). Valoraciones del alumnado universitario sobre las pruebas objetivas. *Revista de Investigación Educativa*, *23*(1), 259-277. https://bit.ly/2wE6z5B

Gilles, J.L., Detroz, P., & Blais, J.G. (2011). An international online survey of the practices and perceptions of higher education professors with respect to the assessment of learning in the classroom. *Assessment & Evaluation in Higher Education, 36*(6), 719-733. https://doi.org/10.1080/02602938.2010.484880

Gipps, C.V. (1994). *Beyond testing: Towards a theory of educational assessment*. Falmer Press.

Goñi-Zabala, J.M. (2005). *El espacio europeo de educación superior, un reto para la universidad*. Ediciones Octaedro.

Goubeaud, K. (2010). How is science learning assessed at the postsecondary level? Assessment and grading practices in college Biology, Chemistry and Physics. *Journal of Science Education and Technology*, *19* (3), 237-245. https://doi.org/10.1007/s10956-009-9196-9

Goubeaud, K., & Yan, W. (2004). Teachers educators teaching methods, assessments, and grading: A comparison of higher education faculty's instructional practices. *The Teacher Educator*, *40* (1), 1-16. https://doi.org/10.1080/08878730409555348

Haladyna, T.M. (1997). *Writing test items to evaluate higher order thinking.* Allyn & Bacon.

Hernández, R. (2012). Does continuous assessment in higher education support student learning? *Higher Education*, *64*(4), 489-502. https://doi.org/10.1007/s10734-012-9506-7

Huba, M.E., & Freed, J.E. (2000). *Learned-centered assessment on college campuses*. Allyn & Bacon.

Hortigüela-Alcalá, D., Pérez-Pueyo, A., & López-Pastor, V. (2015). Implicación y regulación del trabajo del alumnado en los sistemas de evaluación formativa en educación superior. *Revista Electrónica de Investigación y Evaluación Educativa*, *21*(1), ME6. https://doi.org/10.7203/relieve.21.1.5171

Jacobs, L.C., & Chase, C.I. (1992). *Developing and using tests effectively: A guide to faculty.* Jossey-Bass.

Joint Committee on Standards for Educational Evaluation (2015). *Classroom assessment standards: Practices for PreK-12 teachers*. Sage.

Krzykowski, L., & Kinser, K. (2014). Transparency in student learning assessment. *Change*, *46*(3), 67-73. https://doi.org/10.1080/00091383.2014.905428

López-Pastor, V., & Sicilia-Camacho, A. (2016). Formative and shared assessment in higher education. Lessons learned and challenges for the future. *Assessment & Evaluation in Higher Education*, *42* (1), 77–97 http://dx.doi.org/10.1080/02602938.2015.1083535

Mateo, J., & Martínez, F. (2008). *Medición y evaluación educativa*. La Muralla.

McClarty, K.L., & Gaertner, M.N. (2015). *Measuring mastery: Best practices for assessment in competency-based education.* American Enterprise Institute for Public Policy Research. https://bit.ly/2MN1m04

McMillan, J.H. (1997). *Classroom assessment: Principles and practice for effective instruction*. Allyn and Bacon.

McMillan, J.H. (2008). *Assessment essentials for standard-based education* (2da ed.). Corwin Press.

Medina-Díaz, M., & Verdejo-Carrión, A.L. (2019). *Evaluación del aprendizaje estudiantil* (6ta ed.). Autoras.

Medina-Gual, L. (2013). La evaluación en el aula: Reflexiones sobre sus propósitos, validez y confiabilidad. *Revista Electrónica de Investigación Educativa*, *15*(2), 34-50. https://bit.ly/2xm8HiH

Messick, S. (1989). Validity. En R.L. Linn (Ed.), *Educational measurement* (3ra ed., pp. 13-103). Macmillan.

Moss, P.A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice*, *22*(4), 13-25. https://doi.org/10.1111/j.1745-3992.2003.tb00140.x

Muñiz, J., & Fonseca-Pedrero, E. (2008). Construcción de instrumentos de medida para la evaluación universitaria. *Revista de Investigación en Educación*, (5), 13-25. https://bit.ly/39izB8f

Newble, D., & Cannon, R. (1991). *A handbook for teachers in university and colleges.* Kogan Page.

Nitko, A.J., & Brookhart, S.M. (2011). *Educational assessment of students* (6ta ed.). Pearson.

Olmos-Miguelañez, S., & Rodríguez-Conde, M.J. (2010). Diseño del proceso de evaluación de los estudiantes universitarios españoles: ¿Respondiendo a una evaluación por competencias en el espacio europeo y Educación Superior? *Revista Iberoamericana de Educación*, *53*(1). https://bit.ly/2JsjM4B

Palomba, C.A., & Banta, T.W. (1999). *Assessment essentials: Planning, implementing, and improving assessment in higher education*. Jossey-Bass.

Pereira, D.R., & Flores, M.A. (2016). Conceptions and practices of assessment in higher education: A study of Portuguese university teachers. *Revista Iberoamericana de Evaluación Educativa*, *9*(1), 9-29. http://dx.doi.org/10.15366/riee2016.9.1.001

Popham, W.J. (2005). *Classroom assessment: What teachers need to know* (4ta ed.). Pearson.

Poskanzer, S.G. (2002). *Higher education law: The faculty*. John Hopkins University Press.

Pozo, J.I., & Pérez-Echevarría, M. del R. (2009). (Coords.), *Psicología del aprendizaje universitario: La formación de competencias*. Morata.

Rawlusyk, P.E. (2018). Assessment in higher education and student learning. *Journal of Instructional Pedagogies*, *21*, 1-34. https://bit.ly/2vPtbzT

Reddy, Y.M., & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*(4), 435-448. https://doi.org/10.1080/02602930902862859

Ricoy, M.C., & Fernández-Rodríguez, J. (2013). La percepción que tienen los estudiantes universitarios sobre la evaluación: Un estudio de caso. *Educación XXI*, *16*(2), 321-342. https://doi.org/10.5944/educxx1.2.16.10344.

Selke, M.J.G. (2013). *Rubric assessment goes to college.* Rowman & Littlefield.

Smith, J.K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and Practice*, *22*(4), 26-33. https://doi.org/10.1111/j.1745-3992.2003.tb00141.x

Stemler, S.E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*, Article 4. https://doi.org/10.7275/96jp-xz07

Stiggins, R. (2017). *The perfect assessment system*. Association for Supervision and Curriculum Development.

Suskie, L. (2009). *Assessing student learning: A common sense guide* (2da ed.). Jossey-Bass.

Taylor, C.S., & Nolen, S. B. (2005). *Classroom assessment: Supporting teaching and learning in real classrooms*. Pearson.

UNESCO (1998). *La educación superior en el siglo XXI: Visión y acción*. Conferencia mundial sobre la Educación Superior, Paris. https://bit.ly/33Qb4GG

Van der Schaaf, M., Baartman, L., & Prins, F. (2012). Exploring the role of assessment criteria during teachers' collaborative judgement processes of students' portfolios. *Assessment & Evaluation in Higher Education, 37*(7), 847-860. https://doi.org/10.1080/02602938.2011.576312

Voorhees, R.A. (2001). Competency-based learning models: A necessary future. *New Directions for Institutional Research*, *110*, 5-13. https://doi.org/10.1002/ir.7

Webber, K. (2012). The use of learner-centered assessment in US Colleges and Universities. *Research in Higher Education*, *53*(2), 201-228. https://doi.org/10.1007/s11162-0119245-0.

Weimer, M. (Ed.). (2013). *Grading strategies for the college classroom: A collection of articles for faculty.* Magna Publications.

Wolf, K., Dunlap, J., & Stevens, E. (2012). Ten things every professor should know about assessment. *The Journal of Effective Teaching*, *12*(2), 65-79. https://bit.ly/2UBD6Bh

Yükselii, H.S., & Gündüz, N. (2017). Formative and summative assessment in higher education: Opinions and practices of instructors. *European Journal of Education Studies, 3*(8), 336-356. https://doi.org/10.5281/zenodo.832999