The aim of this section is to expand and accelerate advances in methods of teaching bioethics.

Using a Scoring Rubric to Assess the Writing of Bioethics Students

HUGH A. STODDARD, CORY A. LABRECQUE, and TOBY SCHONFELD

Abstract: Educators in bioethics have struggled to find valid and reliable assessments that transcend the "reproduction of knowledge" to target more important skill sets. This manuscript reports on the process of developing and grading a minimal-competence comprehensive examination in a bioethics master's degree program. We describe educational theory and practice for the creation and deployment of scoring rubrics for high-stakes performance assessments that reduce scoring inconsistencies. The rubric development process can also benefit the program by building consensus among stakeholders regarding program goals and student outcomes.

We describe the Structure of the Observed Learning Outcome taxonomy as a mechanism for rubric design and provide an example of how we applied that taxonomy to define pass/fail cut scores. Details about domains of assessment and writing descriptors of performance are also presented. Despite the laborious work required to create a scoring rubric, we found the effort to be worthwhile for our program.

Keywords: bioethics education; assessment of students; performance assessment; high-stakes exam; rubric scoring

Disagreements have been pervasive in educational literature about the types of assessments used to measure learning and the resultant consequences of performance on those assessments.[1,2] Bioethics, as an academic discipline, is not immune to these challenges. Indeed, given the diversity of approaches to teaching the material, reliably assessing whether or not learners have developed problem-solving and professional skills continues to stir debate. As educators, our task is to ensure that our assessments transcend the simple "reproduction of knowledge" and instead target the enhanced skill sets that are meaningful for this interdisciplinary field.[3,4] Yet this goal has remained elusive.

We have embraced these challenges by crafting a comprehensive examination for our master's students that measures a minimal level of competence in learners.[5] By awarding a passing score on the exam, we are certifying that students who are successful on the exam have mastered a subset of the program's competencies. Yet in order for this to be true, faculty members needed to agree on the extent to which the exam's questions—and students' answers—accurately represented these competencies. What counts as a

At the time of the writing of this manuscript, Dr. Schonfeld was an employee of the Emory Center for Ethics.

Cambridge Quarterly of Healthcare Ethics (2016), **25**, 301–311.

doi:10.1017/S0963180115000602

"satisfactory" answer? How much description is enough? What are the characteristics of an "excellent" answer? In order to maximize agreement among faculty graders and minimize interrater variability in exam assessments, we decided to develop a scoring rubric for each exam question.

In this article, we describe the educational theory and practice that supported rubric development, as well as providing a framework for constructing and using rubrics for high-stakes performance assessments, like a comprehensive exam, along with an example rubric for use in a bioethics educational program.

Performance Assessments and Scoring Rubrics

Performance assessments have become more common at all levels of education over the past two decades.[6,7] Performance assessments allow for learners to demonstrate higher-order thinking skills such as applying knowledge in context and making reasoned judgments from core principles.[8] In the case of bioethics education, the desired performance is predominantly in the cognitive and affective domains: students must demonstrate their ability to identify, assess, and address ethical issues in a broad array of situations, as well as reflecting important foundational and theoretical material in support of well-crafted, carefully articulated arguments. Given this, the most effective means of assessing bioethics master's degree students is to use written, constructed-response essays.[9] For current purposes, "effective" refers to the notion of authentically representing the nature of the field while maintaining assessment integrity.[10]

Constructed-response assessment formats, such as reports, essays, and theses, are commonly used at all levels of education; however, inconsistency of scoring[11,12] may compel educational leaders to eschew these techniques. Common difficulties encountered in scoring constructed responses include disagreements about degrees of incorrectness in an answer and difficulties in segregating the content of a response from the writing style or grammatical presentation of a response.[13]

One solution, which allows for implementing an appropriate assessment while still maintaining a reliable and valid scoring of results, is to create and apply a scoring rubric that helps faculty assessors to characterize students' written work.[14] We propose that creating and using a scoring rubric to assist in grading essays, particularly on high-stakes exams, can reduce the scoring inconsistencies.

The Benefits of Using a Scoring Rubric

For students, the benefits of using a rubric reside primarily in knowing the faculty's expectations prior to taking an exam. Having the rubric in advance allows students to focus their attention on the skills that they will be expected to demonstrate to complete the exam successfully. This knowledge should guide students to focus on key points during their preparation and encourage them to integrate the knowledge they have learned over the multiple courses taken in the program. After an exam, a completed rubric that was used to assign a score to a student's work provides specific feedback about his or her performance. This is particularly valuable for aspects in which improvement is necessary, but it also reinforces that successful students have mastered the program outcomes.

Faculty will benefit from increased consistency in scoring of student work by having a standard according to which to measure students' responses that has been crafted explicitly to reflect the

expectations of each answer. This will minimize interrater variability and therefore minimize the influence of a particular faculty member's interests or biases. For this reason, students may perceive that the use of a scoring rubric is fairer than alternative methods of correction. Additionally, scoring rubrics benefit faculty by generating a standard by which program quality can be evaluated. A key point for educational program evaluation is the performance of its graduates. Using scoring rubrics to produce a consistent measure of student performance provides a valuable longitudinal perspective on the program.

Example of Rubric Development for Bioethics

For a high-stakes, comprehensive exam option used in one bioethics master's degree program,[15] we developed a rubric that was adapted from a learning framework that had been originally proposed by John Biggs.[16] This framework, called the Structure of the Observed Learning Outcome (SOLO) taxonomy, is very well suited for the population and educational level of bioethics students. The SOLO taxonomy postulates five levels of learning outcomes, ordered by their complexity.[17,18] Because the increasing levels of complexity closely matched the levels of performance that we expected for bioethics students, we adopted this taxonomy as the theoretical framework for our rubric (see Table 1).

The SOLO framework[19] includes the following descriptive categories:

1) *Prestructural*: The task is not attacked appropriately; the student has not understood the point.
2) *Unistructural*: One or a few aspects of the task are picked up and used (understanding as nominal).
3) *Multistructural*: Several aspects of the task are learned but are treated separately (understanding as knowing about).
4) *Relational*: The components are integrated into a coherent whole, with each part contributing to the overall meaning (understanding as appreciating relationships).
5) *Extended abstract*: The integrated whole at the relational level is reconceptualized at a higher level of abstraction, which enables generalization to a new topic or area, or is turned reflexively on oneself (understanding as far transfer, and as involving metacognition).

Sample Bioethics Question

To demonstrate how a rubric would be constructed in bioethics, consider the following question, which might appear on a master's-level bioethics examination. Note that four "domains" are marked in the question for future reference.

> Defend the moral permissibility of abortion from the perspective of Mary Ann Warren (Domain 1) and Peter Singer (Domain 2). Then critique the moral permissibility of abortion as justified by personhood criteria (Domain 3). What are the implications of the fact that personhood can be used to justify both the moral permissibility of abortion and the moral impermissibility of abortion (Domain 4)?

Throughout the remainder of this manuscript, we will refer to this question in order to demonstrate how to apply the theoretical principles of rubric development and interpretation to bioethics education.

General Principles of Creating a Scoring Rubric

There are various designs for scoring rubrics, but we focus only on a common

Table 1. Matrix Rubric for the Bioethics Example

	Prestructural	Unistructural	Multistructural	Relational	Extended abstract
Defense 1 (Warren)	Descriptor	Descriptor	Descriptor	Descriptor	Descriptor
Defense 2 (Singer)	Descriptor	Descriptor	Descriptor	Descriptor	Descriptor
Critique	Descriptor	Descriptor	Descriptor	Descriptor	Descriptor
Implication	Descriptor	Descriptor	Descriptor	Descriptor	Descriptor

Note. The far left-hand column lists domains for assessment. The top row lists categories of student performance. Descriptors of performance for each level of each domain are shown in the remaining cells.

matrix-style grid that exemplifies the key features and benefits of using a rubric. The grid uses horizontal and vertical axes with one construct shown on each axis. Cells on the grid represent the intersection of those two constructs. "Descriptors" are text descriptions of student work that would be characteristic of that point of intersection of the horizontal and vertical constructs.

When designing a matrix rubric, the matrix is commonly laid out with the criteria (i.e., assessment domains) listed on the vertical axis and the levels of performance on the horizontal axis. The criteria may be listed in any order; however, the levels should be ordinal by increasing performance quality. For our example, the criteria are as follows: Defense 1 (Warren), Defense 2 (Singer), critique, and implication. The horizontal axis follows the SOLO categories[20] from left to right: prestructural, unistructural, multistructural, relational, and extended abstract. Each cell formed by the intersections of the rows and columns should contain a short description of the characteristics of an answer related to that criterion (row) and of that quality (column).

The decisions that need to be made during rubric development are as follows: What criteria should be used in the assessment domains? How many performance levels should be specified? What are the defining characteristics for each performance level in each domain, and how will those characteristics be stated (i.e., how should the descriptors be worded)? Careful consideration of these three decisions is critical in order to create a rubric that achieves its goals of producing consistent scores for student work and that reinforces the learning outcomes of the educational program.

Rubric Domains

The assessment domains should be discrete, so that as raters make judgments about various attributes of students' work, they are clear about in which domain those attributes should be reflected. For our purposes, the easiest way to accomplish this was to ensure that both students and faculty were clear about the topics students should include in a thorough response. We achieved this by adding parenthetical references to indicate the discrete domains of the question, as marked in the preceding sample question.

Based on the principles of structure described previously and the sample question we posed, the assessment domains in our rubric would be as stated in the example in the previous section. Note that each part of the question represents a unique domain for which a performance level must be assigned. Identifying the relevant domains is significant work. In our case, the process helped us to re-evaluate what we were truly asking of the students and often prompted us to rework the question

itself. In this way, rubric creation is often bidirectional—careful attention to scoring requires careful attention to wording the question in a manner that minimizes misinterpretations by students or faculty.

The number of assessment domains in the rubric depends on the purpose of the assessment. For example, a holistic rubric with a single domain may serve well when the assessment is intended to provide a global assessment of the student. If the assessment's goal is to provide students with feedback on strengths and weaknesses, then an analytic rubric with a larger number of domains is appropriate. As a guide, analytic rubrics typically employ three to five discrete domains. Aggregating scores from multiple domains into a single grade is discussed later in this article.

Levels of Performance

To develop a rubric for assessment in bioethics education, the SOLO taxonomy can straightforwardly be applied to each of the domains being assessed. For example, a student who demonstrates knowledge of some bioethics concepts would be at the unistructural level. A student who knows basic concepts and how those concepts can be applied in various cases would be at the multistructural level. If a student can consider a case and explain how she or he would apply bioethics concepts and reasoning to the case and can also demonstrate how those concepts connect to each other, then that would be at the relational level of performance. The extended abstract level would be achieved only by a student who can extend concepts and applications to hypothetical bioethics situations that have not previously been considered, who can recognize the broader implications of her or his claims and reasoning, and who can critically evaluate the validity of her or his own claims and reasoning.

The levels of student performance, shown on the horizontal axis of a scoring rubric, should be ordinal but not necessarily a scale. Whether the lowest level is placed at the left, with increasing levels toward the right, or vice versa, with the highest level at the left, is a matter of preference. There is one major caveat for defining the performance levels: we recommend that the levels not be assigned numeric values. Granted, the levels are ordinal; however, associating a number with a level imposes arithmetic relationships among those levels. For example, a paper rated as being "level 4" is twice as good as one rated as "level 2." The temptation for rubric users to overinterpret numeric results is almost inescapable. The descriptors used in rubrics for bioethics education are composed of qualitative observations. Imposing quantitative principles on qualitative characteristics is not justifiable and leads to confusion or misinterpretation.

Domain-Specific Descriptors of Performance

The descriptors that are written for the cells of a rubric are the core of the rubric's value to educators and students. These descriptors exemplify the traits that should be evident at each performance level within each domain. In order to realize their full value, descriptors must be specific and evident and not contain judgmental terms such as "good" or "excellent." (It should be emphasized here that overall evaluation of students' performance, when judgments such as "good" or "excellent" are appropriate, ensues after measuring students' abilities using the scoring rubric.) Descriptors should be concise and should identify observable features of student performance. These features

should be stated in precise language such that raters can readily find evidence of the features in the student work—or can note the absence thereof. The reliability of a rubric as an assessment tool depends on the consistency with which raters score student work. Although training the raters might increase reliability, we contend that well-written descriptors are vital to ensuring that raters draw similar conclusions about the quality of students' work.[21,22]

For example purposes, Table 2 demonstrates what descriptors could look like for one of the domains of assessment that we previously introduced. In the same manner, faculty would need to compose descriptors for each of the other assessment domains for the question. Note that in Table 2 we include very specific descriptors throughout, including itemizing the components necessary to constitute a complete answer. Doing so minimizes confusion among graders, because it provides them with an exact list in the rubric of what is expected. An added benefit is that disagreements among faculty members regarding a "correct" answer are negotiated during rubric construction rather than while trying to score students' responses. It is impossible to anticipate all of the ways in which students' answers may deviate from what is expected, but we also highlight both the likely deficiencies and the elements of excellence in students' answers, so that faculty will recognize where an answer belongs on the rubric scale.

Building on the matrix model described previously with criteria listed on the vertical axis and SOLO levels of quality on the horizontal, we propose the following descriptors for the first domain of assessment (Defense 1 [Warren]):

- *Prestructural*: Fails to identify any of the criteria for personhood; confuses Warren's argument with others'.
- *Unistructural*: Recognizes that Warren's theory depends on constitutive traits; identifies two or fewer traits correctly.
- *Multistructural*: Correctly describes at least three traits of Warren's theory: (1) consciousness and the capacity to feel pain, (2) reasoning, (3) self-motivated activity, (4) the capacity to communicate, and (5) the presence of self-concept and self-awareness.
- *Relational*: Describes all five traits of Warren's theory and describes the relationships among those traits.
- *Extended abstract*: Discusses contemporary implications of Warren's theory (such as findings about consciousness) and presents hypothetical situations to test applications of that theory.

As mentioned previously, although we do not provide detailed examples here, descriptors for each of the other criteria and levels of performance need to be composed and agreed on by the faculty.

We readily recognize that some readers may disagree with the domains or descriptors in our rubric. In the opinion of some, we may have failed to include the most important features of an answer, or we may have included too much, or we set the descriptors at the "wrong" performance level. Yet this is precisely the point of creating a rubric through a consensus process. The only acceptable rubric for any individual or program will be the one that is congruent with the program's curriculum and outcomes. We have provided an example here simply to demonstrate the process of applying a theoretical framework, not to suggest a singular, definitive instrument for use in all contexts.

Score Interpretation

Up to this point, we have presented the process for developing a scoring rubric

https://doi.org/10.1017/S0963180115000602 Published online by Cambridge University Press

Table 2. Bioethics Rubric with Minimum Pass and Student Feedback Information

	Prestructural (FAIL)	Unistructural (FAIL)	Multistructural (PASS)	Relational (PASS)	Extended abstract (PASS)
Defense 1 (Warren)	Fails to identify any of the criteria for personhood; confuses Warren's argument with others'.	Recognizes that Warren's theory depends on constitutive traits; identifies two or fewer traits correctly	Correctly describes at least three traits of Warren's theory: (1) consciousness and the capacity to feel pain, (2) reasoning, (3) self-motivated activity, (4) the capacity to communicate, (5) the presence of self-concept and self-awareness.	Describes all five traits of Warren's theory and describes the relationships among those traits.	Discusses contemporary implications of Warren's theory (such as findings about consciousness) and presents hypothetical situations to test the applications of that theory.
Defense 2 (Singer)	Descriptor A2	Descriptor B2	Descriptor C2	Descriptor D2	Descriptor E2
Critique	Descriptor A3	Descriptor B3	Descriptor C3	Descriptor D3	Descriptor E3
Implication	Descriptor A4	Descriptor B4	Descriptor C4	Descriptor D4	Descriptor E4

Note. The far left-hand column lists domains for assessment. The top row lists categories of student performance. Descriptors of performance for each level of each domain are shown in the remaining cells. The shaded areas represent the categories that we considered to be failing performance.

that will measure student performance in a valid and reliable way. To complete the educational assessment process, student performance must be interpreted and a decision must be made about the students' achievement of desired outcomes. After students' products have been scored using a rubric, those scores need to be interpreted, often using a scale of A-B-C-D-F or simply pass/fail. The interpretation of scores has a significant impact on the students' educational trajectory, which is a compelling reason for dedicating a substantial amount of time and effort to developing a valid and reliable rubric for scoring students' work and interpreting their scores. For example, the descriptors presented in the five levels suggested by the SOLO taxonomy give clear and specific guidance to the reader regarding into which performance level an essay should be categorized.

As stated previously, we advise against assigning numeric values to performance levels on a rubric. Although it is not uncommon for raters to assign numbers using a rubric, convert those numbers into percentages, and then interpret the percentages as letter grades, we strongly advise against this practice. We noted previously that doing so requires the assumption that the descriptors for performance levels represent arithmetic relationships (e.g., half as good, 25% better, etc.), and this assumption is practically impossible for the nonquantitative outcomes of bioethics education. By way of contrast, a rubric used to score a musical performance in which one domain is "intonation" could maintain arithmetic properties for performance levels by using descriptors such as "all notes were in tune" and "90% of notes were in tune."

Score interpretations could be made using common grading scales, such as A–F, or categorically, such as categorizing a student's ability as "emerging," "proficient," or "outstanding." Because our goal was to assess a high-stakes examination, we were most concerned with a pass-or-fail decision. The principles of setting a cut score, described subsequently, may be extrapolated for use in discriminating among other categories as well.

Setting Cut Scores

Interpreting rubric measurements in terms of grades, particularly pass/fail decisions, involves setting a minimum standard for performance.[23] We will refer back to the SOLO taxonomy to provide a theoretical basis for setting minimum standards appropriate to bioethics education. A basic tenet of setting the "cut score" for minimum passing is that the cut point should be congruent with the stated outcomes of the course or program that is being assessed. Because bioethics programs are usually offered at advanced baccalaureate, professional, or graduate school levels, faculty will likely consider minimum acceptable performance to be at the multistructural level or the relational level of the SOLO taxonomy. This is illustrated by the shaded areas on Table 2. The choice of cut points should be made and justified by each program in consideration of the program's outcomes and student population. Techniques and additional considerations are described elsewhere in excellent detail.[24] For our own exam, we set the minimum passing standard at the multistructural level. Thus, a student scoring at the multistructural level would pass, whereas a student at the unistructural level would fail.

The most important element to successfully using a rubric is to write descriptors that precisely state the characteristics that distinguish minimally passing work from failing work. Apart from making that critical distinction, the number and arrangement of descriptors may be varied according to the purposes of the assessment. Thus, there are two

approaches to aligning cut scores with descriptors. On one hand, if the assessment is strictly to identify substandard performance with no concern for discrimination between excellent work and minimally passing work, then only two performance levels are needed, along with the descriptors for those two levels. Specifically, only descriptors for the borderline between the lowest passing level and the highest failing level would need to be written. To create such a rubric from our example, only the titles on the horizontal axis would change. Rather than using all of the SOLO categories as titles, only two columns—failing and minimally passing—would need to be included. Given our passing standard elucidated previously, the rubric would retain only the criteria and descriptors in the unistructural and multistructural columns.

On the other hand, if the faculty has already done the work described previously using the SOLO taxonomy, then they simply need to reach consensus about which category is the lowest acceptable level to pass the exam. An educational advantage of using this full-scale approach is that, in addition to making pass/fail decisions, faculty can provide feedback to the learners about relative areas of strength and weakness. To construct such a rubric, at a minimum, descriptors must be defined for the extremes of each domain as well as for both sides of the cut point between pass and fail. Any additional points on the quality scale could be left for interpolation by the rater.

Aggregating Scores from Multiple Domains

Using analytic rubrics that have multiple domains requires one more step for interpretation than is needed for holistic rubrics. In order to represent a student's overall performance with a single score, such as a grade, the student's performance on the various domains of measurement needs to be aggregated. This can be done using either a compensatory approach or a conjunctive approach. In the compensatory approach, a student can compensate for low performance in one domain with high performance in another. In the conjunctive approach, the student must achieve the minimum standard on each domain in order to pass the overall assessment. The decision on which model to use depends entirely on the purpose of the assessment and the consequences, such as participation in a remedial program, that are dispensed to the students according to their results.

For our own comprehensive exam,[25] students responded to several questions with a separate rubric created for each question. We used a compensatory approach when aggregating the domains of an individual question, so a student who did poorly in one domain could recover with a high score in another domain for that question. But we used a conjunctive approach to arrive at a pass-or-fail decision for the exam as a whole. Thus students needed to achieve a passing score for each question they answered.

Additional Considerations for Rubric Creation

Creating a scoring rubric is a laborious process. Making decisions about the student attributes that should be rewarded by the program will bring to the forefront discrepancies in values among faculty as to what a graduate should know, be, and do. Creating a scoring rubric to be used for high-stakes, program-level assessments forces faculty members to confront their own disagreements and negotiate common ground. In essence, creating the rubric forces faculty members to discuss and

compromise on these differences, differences that otherwise could fester or flare up into departmental infighting. We propose that bringing such disagreements to the forefront, as is necessary to develop a scoring rubric, is not always pleasant but is more collegial and better for students and the program as a whole than compelling students to continually adapt to the varying expectations and standards that may be held by various instructors.

Conclusion

The benefits of using a scoring rubric are congruent with the desired outcomes of bioethics education. The nature of bioethics education demands that assessments reinforce the critical reasoning and applications of knowledge that characterize the discipline. Scoring rubrics provide two major benefits to bioethics educators. First, rubrics increase the reliability of performance assessment scores, such as essay exams or other performance products. This improved consistency ensures fair and accurate grading of students. Second, the process of developing a rubric requires faculty members to negotiate differences of opinion about the outcomes and pedagogy for the program and encourages communication with students about the expected outcomes.

While undergoing this development process ourselves and applying it to a comprehensive exam for students in a bioethics master's program, we found that following the principles of rubric design was time-consuming and frustrating at the moment. But we concluded that building a strong theoretical footing for our rubric was richly rewarded after implementation. In particular, the SOLO taxonomy provided us with a theoretical basis that suited our program goals well and helped us organize our thinking while we defined the performance levels and wrote the descriptors for our rubric. We concluded that our students and the program as a whole have benefited from this effort.

Notes

1. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing, 2014 Edition*. Washington, DC: American Educational Research Association; 2014.
2. Pellegrino JW, Chudowsky N, Glaser R. *Knowing What Students Know: The Science and Design of Educational Assessment*. Washington, DC: National Academies Press; 2001.
3. Antes AL, Murphy ST, Waples EP, Mumford MD, Brown RP, Connelly S, et al. A meta-analysis of ethics instruction effectiveness in the sciences. *Ethics and Behavior* 2009;19(5):379–402.
4. Mumford M, Connelly S, Brown R, Murphy S, Hill J, Antes A, et al. Ethics training for scientists: Effects on ethical decision-making. *Ethics and Behavior* 2008;18(4):315–39.
5. Schonfeld T, Stoddard HA, Labrecque CA. Examining ethics: Developing a comprehensive exam for a bioethics master's program. *Cambridge Quarterly of Healthcare Ethics* 2014;23(4):461–71.
6. Brookhart SM. Assessment theory for college classrooms. *New Directions for Teaching and Learning* 2004;2004(100):5–14.
7. Stiggins RJ. Design and development of performance assessments. *Educational Measurement: Issues and Practice* 1987;6(3):33–42.
8. Lane S, Stone CA. Performance assessment. In: Brennan RL, ed. *Educational Measurement 4th Edition*. Westport, CT: American Council on Education and Praeger; 2006.
9. Favia A, Frank L, Gligorov N, Birnbaum S, Cummins P, Fallar R, et al. A model for the assessment of medical students' competency in medical ethics. *AJOB Primary Research* 2013;4(4):68–83.
10. See note 1, AERA/APA/NCME 2014.
11. Lohfeld L, Goldie J, Schwartz L, Eva K, Cotton P, Morrison J, et al. Testing the validity of a scenario-based questionnaire to assess the ethical sensitivity of undergraduate medical students. *Medical Teacher* 2012;34(8):635–42.
12. Tierney R, Simon M. What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation* 2004;9(2):1–10 [cited 26 Dec 2013]; available at http://pareonline.net/getvn.asp?v=9&n=2 (last accessed 9 Sept 2014).

13. Lukhele R, Thissen D, Wainer H. On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement* 1994;31(3):234–50.
14. Moskal BM, Leydens JA. Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation* 2000;7(10): 71–81 [cited 26 Dec 2013]; available at http://pareonline.net/getvn.asp?v=7&n=10 (last accessed 9 Sept 2014).
15. See note 5, Schonfeld et al. 2014.
16. Biggs JB, Collis KF. *Evaluating the Quality of Learning*. New York: Academic Press; 1982.
17. See note 16, Biggs, Collis 1982.
18. Biggs J. John Biggs: Writer, academic, traveller; 2013 [cited 26 Dec 2013]; available at http://www.johnbiggs.com.au/academic/solotaxonomy/ (last accessed 9 Sept 2014).
19. See note 16, Biggs, Collis 1982.
20. See note 16, Biggs, Collis 1982.
21. See note 6, Brookhart 2004.
22. See note 14, Moskal, Leydens 2000.
23. Hambleton RK, Pitoniak MJ. Setting performance standards. In: Brennan RL, ed. *Educational Measurement 4th Edition*. Westport, CT: American Council on Education and Praeger; 2006.
24. See note 23, Hambleton, Pitoniak 2006
25. See note 5, Schonfeld et al. 2014.