REVIEW

Strategies for Measuring Advances in Pharmacy Student Learning

Ashley N. Castleberry, PharmD, MEd,a Melissa S. Medina, EdD,b,e Adam Persky, PhD,c,e Lauren S. Schlesselman, PharmD, MA Ed Psychd,e

a University of Texas at Austin, College of Pharmacy, Austin, Texas
b University of Oklahoma, Health Sciences Center College of Pharmacy, Oklahoma City, Oklahoma
c University of North Carolina, Eshelman School of Pharmacy, Chapel Hill, North Carolina
d University of Connecticut, Center for Excellence in Teaching and Learning, Storrs, Connecticut
e Associate Editor, American Journal of Pharmaceutical Education, Arlington, Virginia

Submitted April 28, 2021; accepted August 3, 2021; published April 2022.

Objective. To provide guidance to authors and reviewers on how to design and evaluate educational research studies to better capture evidence of pharmacy student learning.

Findings. A wide variety of assessment tools are available to measure student learning associated with educational innovations. Each assessment tool is associated with different advantages and disadvantages that must be weighed to determine the appropriateness of the tool for each situation. Additionally, the educational research design must be aligned with the intent of the study to strengthen its impact.

Summary. By selecting research methods aligned with Kirkpatrick’s levels of training evaluation, researchers can create stronger evidence of student learning when evaluating the effectiveness of teaching innovations.

Keywords: student learning, educational research, assessment of learning

INTRODUCTION

Educators conduct educational research to inform the Academy on how to improve learning for future generations. The more instructors understand about student learning, the more they can do to improve it. With that charge, we hope to make evidence-based decisions on how to get students to learn the desired knowledge and skills. Ultimately, we hope the knowledge and skills translate to workplace behaviors and positive patient outcomes. However, educators’ ability to make decisions on what is best and in what context depends on their ability to think critically and to have quality evidence to think about. It is the latter that concerns the educational community: do we have quality evidence, and if not, how do we generate it? The objective of this review was to provide guidance to educators on how to design educational research studies to better capture evidence of student learning.

To frame this article, we began with a discussion of taxonomies for training and/or evaluating the level of evidence in the scholarship of teaching and learning (SOTL) or an educational research study. Educators should keep in mind that evaluating the level of evidence in an educational research study is akin to assessing evidence in clinical practice. In clinical practice, the opinions of authorities or experts are considered the lowest type of evidence, while the systematic reviews and meta-analysis of relevant randomized controlled trials are considered the highest. For instance, to answer a clinical question about the best therapy, the educator should refer to randomized controlled trials, meta-analysis, and to a lesser extent, cohort studies, case-control studies, and cases series, remembering that the outcomes are related to the actual goals of the intervention.

One popular taxonomy is Kirkpatrick’s levels of training evaluation. At its base are feelings or perceptions. This level captures how the learner feels or perceives the training has or has not benefited them. This is considered low level evidence because perceptions are biased and, in general, humans are not great judges of what they know or can do. For example, according to the ease of fluency bias, people feel that things that are learned easily are well learned and things that come to mind quickly are correct. This is a bias. Just because something seems easy, does not mean it is learned, and answers that come to mind first are not necessarily the correct ones. The main purpose of the perceptions level is to obtain positive comments about how the learners viewed the program and
whether they would attend future training programs, and to obtain negative comments to improve the training program.7 The comments themselves do not guarantee learning has occurred.

Above the level of perception is knowledge and conceptual knowledge, which includes the more traditional “paper and pen” facts. This type of assessment still has some subjectivity as the trainer selects the items to be tested; nevertheless, the responses are less subjective than when trainees are asked only about their perceptions. This level of assessment establishes that students can recall facts and perform skills when prompted. It is these facts and skills that are required to enact behaviors. As stated in some of the original work by Kirkpatrick, without learning, as measured by knowledge acquisition, no behavioral change will occur.7

In fact, behavioral change is the third level of the taxonomy. At this level, the instructor asks the learner to perform behaviors they have been trained to do and assesses whether the learner’s performance improved as a result of the training. Can they execute the skill and teach others the skill? The ability to perform a skill and to teach the skill to others, such as a patient, requires a higher level of understanding than can be assessed by a simple knowledge test. Ultimately, the goal of training is to improve student outcomes and organizational performance, along with patient outcomes in the case of health education; thus, the third level is quite important.

The last level is results. While behaviors are great, the educator needs to know whether the behaviors translate to positive results, just like they hope innovation leads to better outcomes and they are not “innovating for innovation’s sake.” Thus, the results of the behavior are the highest level of training outcomes. The hard part is determining the outcome. For example, at a university, is the desired outcome improved graduation rates for learners, the number of students accepted to postgraduate training positions or jobs, or decreased costs of training for the learner? This fifth level is the return on investment or cost-effectiveness.7 That is, does the training deliver results and at what cost? This is an important issue in educational research: what strategies result in better learning, in what contexts, and at what cost?

While this model can be useful for evaluating the level of evidence used in a research study, Kirkpatrick’s model has its limitations. The first is its hierarchical nature, ie, higher levels are more important. This hierarchy can be problematic because all levels are required for decision making. If a training program leads to behavior change but the learners did not perceive the experience well, that may impact what changes should be considered for future offerings of the program. Another limitation is the intercorrelations between levels; the knowledge needed for behavioral change is also needed for results. There are no guarantees that improved perceptions lead to more knowledge or that more knowledge leads to behavioral change or that behavioral change leads to results. The Kirkpatrick model can still be a useful framework for evaluating levels of evidence if researchers address these limitations. However, several other models have been developed that researchers can use, such as the input-process-output (IPO) model; Brinkerhoff’s six-stage model; Context, Input, Process and Product (CIPP) model; Context, Input, Reaction, and Outcome (CIRO) model; and Kaufman and Keller’s five levels of evaluation.7-9 Overall, whatever model(s) a researcher selects to evaluate their evidence, they should remember that the goal of educational research is to find what benefits student learning and in what contexts. Therefore, they need to establish that learning has occurred and that it can be linked to an intervention. To guide researchers in this evaluation process, the remainder of this review will present strategies faculty can use to assess student learning through the lens of various models, including Kirkpatrick’s model as it offers clear steps to follow to gain insights about learning outcomes. Faculty can then use this guidance to better evaluate and document student changes. If disseminated, this will lead to improvement in the quality of educational literature.

DISCUSSION

Many tools and measures are used to assess educational interventions in pharmacy education. These can be categorized based on Kirkpatrick’s taxonomy of training evaluation. The first level, reaction, includes measures such as learner perceptions of confidence or satisfaction in the experience. These are commonly evaluated with self-assessments or course evaluations. The second level, learning, is often assessed using tools like quizzes (planned or surprise), examinations, course grades, and even the Pharmacy Curriculum Outcomes Assessment (PCOA). Behaviors, level three, can be evaluated using objective structured clinical examinations (OSCEs) and experiential learning, both introductory pharmacy practice experiences (IPPEs) and advanced pharmacy practice experiences (APPEs). The highest level, results, can be determined through assessment of Entrustable Professional Activities, or by results on the North American Pharmacist Licensure Examination (NAPLEX) and Multistate Pharmacy Jurisprudence Examination (MPJE), as well as by performance in professional practice after graduation. Each measure has benefits and limitations that researchers must carefully consider before implementing them (Appendix 1 and Table 1).
The first aspect to consider is the research question. The research question should be meaningful, clear, and relevant to the advancement of pharmacy education and educational theory. The nature of the research question will determine the best approach to use. Cook and colleagues10 developed a hierarchy for medical education research based on the intent of the study. The three main categories are description, justification, and clarification. The lowest level, description, refers to studies that present an innovation for which there is no available comparison. Justification, the middle category, refers to studies that compare the effectiveness of different educational interventions to determine which is best. Clarification, the top category, offers advancement for the educational literature because these studies answer the questions of how and why the intervention works. Strongly aligning the educational research design with the intent of the study strengthens the impact of the findings on others in the Academy.

Most educational research evaluates the impact of an intervention on student learning. Although research can evaluate why and how this impact occurred through design-based research, to date this is only undertaken in a limited number of studies.

Table 1. Solutions to Common Research Problems

<table>
<thead>
<tr>
<th>Methodology</th>
<th>Problem</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessments</td>
<td>Pre-Post</td>
<td>Only measuring before and after an intervention does not capture long-term retention.</td>
</tr>
<tr>
<td>Post only</td>
<td>Not capturing participants at baseline makes it difficult to evaluate improvement as a result of the intervention.</td>
<td>Add a baseline assessment.</td>
</tr>
<tr>
<td>Pre/Post/Long-term</td>
<td>May lose participants to attrition especially if the follow-up is 1 year later or after participants graduate.</td>
<td>Ensure you tell participants about the plans for follow-up and ask them to provide current contact information. Consider shortening the long-term time window, if possible, to increase access to participants.</td>
</tr>
<tr>
<td>Tools</td>
<td>Surveys, validated</td>
<td>Validation is required. If you change them, you have to validate them again.</td>
</tr>
<tr>
<td>Rubrics55</td>
<td>Rubrics may not be able to assess every research construct. Assessors may interpret rubric differently. Rubric may not be validated.</td>
<td>Gain input from multiple sources on rubric creation. Train assessors on how to use the rubric to increase reliability. Validate rubric using different evaluators or institutions. Review and report statistics regarding rubric performance.</td>
</tr>
<tr>
<td>Psychological instruments</td>
<td>Instrument may not fit study objectives. Cost can be prohibitive.</td>
<td>Contact authors for collaboration and possible discounts. Explore options for grant funding.</td>
</tr>
<tr>
<td>Models</td>
<td>Cross Over/Within Subjects41</td>
<td>Challenging to do in a classroom. Topic difficulty is a potential confounder.</td>
</tr>
<tr>
<td>Between Subjects</td>
<td>Can be difficult to ensure that the two cohorts are equivalent May take a long time to collect data if using the same class in different years.</td>
<td>Assess baseline skills to know if there are substantial differences in baseline knowledge or skills. Could use different sections of the same course, if available.</td>
</tr>
</tbody>
</table>
A well-designed research project should model the backward design of the learning model. As outlined by Wiggins and McTighe,11 backward design of learning experiences starts by identifying objectives, ie, what students should know, be able to do, or believe by the end of the learning cycle. This is followed by creation of the assessment to measure changes in what students know, can do, or believe. The final step is planning the sequence of lessons and activities that will prepare students to successfully complete the assessment. In contrast, weaker research projects (and educational experiences) identify a topic or content that needs to be covered, plan lessons and activities to teach that topic, then create an assessment to measure learning or simply evaluate student perception of the experience of learning.

Evaluation of the impact of the intervention on students' knowledge, skills, and attitudes enables researchers to analyze results at Kirkpatrick's second level and above. However, to evaluate the impact of the intervention on achievement of learning objectives, researchers must first define the measurable learning objectives for the course or activity. There are several tiered taxonomies of learning to use to accomplish this (Table 2).

Bloom's taxonomy of educational objectives remains the most widely known taxonomy for developing learning objectives.12 This hierarchical ordering of cognitive learning consists of six levels ranging from remembering to creating, with the upper three levels requiring learners to use higher-order thinking skills. At the lowest level of the taxonomy (remembering), learners are recognizing and recalling facts. As they move up the hierarchy, learners start to understand the meaning of the facts (understanding) and to apply the facts, rules, concepts, and ideas (applying). From here, learners begin to break down information into component parts (analyzing), judge the value of information or ideas (evaluating), and finally combine parts to make a new whole (creating).

Fink’s taxonomy of significant learning takes an interactive rather than hierarchical approach to learning.13 According to Fink’s taxonomy, each type of learning can stimulate other types of learning to occur. The six levels of Fink’s taxonomy are foundational knowledge, application, integration, human dimension, caring, and learning how to learn. Foundational knowledge, serving as the base of learning, involves the learner’s ability to remember and understand information. The application of knowledge occurs when learners develop critical-, creative-, and practice-thinking abilities. Integration represents the making of connections between knowledge, ideas, perspectives, and learning experiences. Human dimensions of learning involve learning about self (personal) and others (social), particularly how to use reflection and feedback to identify areas of strength or that need improvement, along with how to interact with others. Caring requires the development of feelings about something new or caring in a new way. Finally, learning how to learn involves becoming a better student and a self-directed learner so that learning continues beyond the learning experience.

When developing learning objectives related to the affective domain, Krathwohl’s taxonomy is one of the best known.14 The five-tiered taxonomy is ordered according to the principle of internalization, a process in which a learner’s affect ranges from general awareness to acting consistently in accordance with a set of values. At the lowest level (receiving), learners become aware of or sensitive to the existence of ideas and phenomena, along with being willing to tolerate them. At the responding level, learners go beyond attending to a phenomenon to react to it in some way, including questioning new ideals, concepts, and models to fully understand them. When learners reach the valuing level, they are concerned with the worth or value they attach to a phenomenon, behavior, or object, ranging from simple acceptance of a value to a more complex level of commitment. The organization level is concerned with

Table 2. Different Types of Available Hierarchical Models Across Knowledge, Skills, and Attitudes

<table>
<thead>
<tr>
<th>Cognitive Domain (Knowledge)</th>
<th>Affective Domain (Attitudes)</th>
<th>Psychomotor Domain (Skills)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bloom’s Taxonomy</strong></td>
<td><strong>Fink’s Taxonomy</strong></td>
<td><strong>Krathwohl’s Taxonomy</strong></td>
</tr>
<tr>
<td>Creating</td>
<td>Learning how to learn</td>
<td>Characterization of a value</td>
</tr>
<tr>
<td>Evaluating</td>
<td>Caring</td>
<td>Organization</td>
</tr>
<tr>
<td>Analyzing</td>
<td>Human dimension</td>
<td>Valuing</td>
</tr>
<tr>
<td>Applying</td>
<td>Integration</td>
<td>Responding</td>
</tr>
<tr>
<td>Understanding</td>
<td>Application</td>
<td>Receiving</td>
</tr>
<tr>
<td>Remembering</td>
<td>Foundational knowledge</td>
<td></td>
</tr>
</tbody>
</table>

341
Selecting Assessments

Just as assessment of learning in the classroom must be matched with the intended learning objectives, so must the tool used in a research project be matched with learning objectives when evaluating whether the educational intervention impacted learning. A study by FitzPatrick and colleagues reported poor content validity of assessments because learning objectives and assessments were mismatched or insufficient assessment tasks were conducted to adequately assess student learning. To sufficiently assess learning, each objective should have at least five to six tasks aligned with it.

A handful of processes are available to determine alignment between learning objectives or standards and assessments, including the Webb alignment process. The Webb process compares standards and assessments on depth-of-knowledge consistency, categorical concurrence, range-of-knowledge consistency, and balance of representation. While this process may be more in depth than most educational research projects warrant, it does provide a solid framework. Assessment tasks should not only cover the topic but be at the learning objective’s intended level of a learning taxonomy or as required in pharmacy practice.

Each item in an assessment tool should be clearly aligned with one or more objectives with the alignment shared here. Multiple assessment items should be aligned to each learning objective. To demonstrate a change in knowledge, skills, and attitudes, researchers should provide data on the change in learning for each of those learning objectives.

In the classroom, student perceptions of learning or of the educational experience are not the measure used to evaluate their knowledge, skills, or attitudes. Similarly, student perceptions should not be the only measure used to evaluate the success of a learning intervention. As outlined above, reactions, such as satisfaction, and changes in attitude are Kirkpatrick’s lowest levels for analyzing and evaluating the results of an educational experience.

In general, individuals tend to be overconfident when evaluating their own performance. This phenomenon, referred to as illusory superiority, was demonstrated by Dunning and Kruger who found that less competent individuals lack metacognition and are unable to recognize their own incompetence, leading them to overrate their ability. Pennycock and colleagues reported that individuals with the greatest number of errors on a cognitive reflection test had overestimated their performance by a factor of more than three. In an evaluation of metacognitive accuracy relative to professional skills, Zell and Kri found a weak correlation between perceived ability and actual ability. A 2020 study found a lack of correlation between student perceived achievement and change in ability.

Despite verification of the Dunning-Kruger effect, many educational researchers’ base claims of effectiveness of an educational intervention solely on student perceptions of learning or attitudes. To better demonstrate achievement, researchers should triangulate using other measures of achievement.

Triangulation is sometimes used to refer to all instances in which two or more research methods are used, but it is also a mixed method approach referring to the use of more than one approach to enhance confidence in the findings and add to the richness and complexity of the inquiry. The concept was first proposed in 1959 by Campbell and Fiske, who advocated for the use of multiple measures to determine the extent to which different measures converged. Denzin and Lincoln argued that triangulation added rigor, breadth, and depth to an investigation. Denzin identified four basic types of triangulation: data, investigator, theory, and methodological. Data triangulation entails gathering data at different times but also with a variety of groups, while theoretical refers to the use of more than one theoretical position in interpreting data. Methodological triangulation refers to the use of more than one method for gathering data; investigator triangulation entails the use of more than one researcher to gather and interpret data.

Pre-post assessments are another method that strengthens research findings. Pre-post assessments are designed to assess learning over a predetermined period of time before and after the learning intervention. The pretest component serves as a baseline assessment of knowledge and skills, while the posttest measures progress resulting from the educational experience.

The pre-assessment, or background knowledge probe as it is described by Angelo and Cross, provides the instructor with insight into student knowledge retained
from previous courses or experiences. The results of the pretest serve two functions. The first is allowing instructors to evaluate whether students possess the prerequisite knowledge, thereby allowing the instructor to adjust the educational intervention to account for this. The second is reducing the assumption that a student’s achievement following the educational intervention is solely the result of the intervention. To achieve these functions, the pretest should contain questions pertaining to essential prerequisite knowledge in addition to questions aligned with the knowledge and skills expected to be gained during the educational intervention.

As with all assessments, items on pre- and post-assessments should be aligned to the learning objectives. This alignment should be made clear in the manuscript so readers know how much of an improvement occurred in response to the educational intervention for each learning objective. The use of pre-post assessments allows for a variety of in-depth statistical analyses beyond change in grade. In addition to paired *t* tests, researchers may consider using the Rasch model fit, test reliability, item response analysis, mean person ability, and other analyses.

A pre-post design can still be limited in providing an accurate understanding of whether the change observed is only attributable to the intervention or how the impact was achieved. They also do not demonstrate sustained change over time or whether students retain the outcomes they gained over time.

Comparison of outcomes for two different groups is another approach for determining the success of an educational intervention. A control group provides an estimate of what the experimental group would have learned or achieved had it not received the educational intervention. The control group ensures the internal validity of the research. When using a control group, the researcher should ensure the same educational outcomes are intended and measured.

Control group comparisons may be equivalent or non-equivalent groups designs. For an equivalent group study, subjects are randomly assigned to either the control group or the experimental group. Each group, chosen and assigned at random, is presented with either the new educational intervention or the standard experience. The control group ensures the internal validity of the research. When using a control group, the researcher should ensure the same educational outcomes are intended and measured.

Utilizing cohorts within the same class can lead to IRB challenges if the control group is not provided with an experience that is superior or is not taught a new topic. To overcome this, researchers can include in the study design and IRB submission a plan to provide the control group with the educational intervention or topic after the study is completed. This can lead to additional learning time for students and uneven assessment requirements, along with teaching challenges. Because of these and other challenges, this study design is less frequently used.

Evaluation of long-term retention of learning, rather than short-term performance, creates a stronger understanding of outcomes. Research on the long-term impact on behavior or patient care provides evaluation at the third and fourth levels of Kirkpatrick’s model when a long-term evaluation of impact is added to the research design. In the rush to publish, many researchers do not design projects evaluating the long-term impact of an educational intervention. Yet, this approach can demonstrate a deeper impact of an educational intervention. This design can assess not only the impact over time but also the relevance to practice by measuring how this learning translates into pharmacist behavior and patient outcomes. Evaluation at these levels can involve repeated assessment at later points in the curriculum, evaluation of behavior during APPEs, or even evaluation postgraduation.

Mixed method research, another means to improve reliability and validity of results, involves a combination of qualitative and quantitative dimensions “for the broad purposes of breadth and depth of understanding and corroboration.”26 The goal of mixed methods is to expand the knowledge obtained and validate the findings. To achieve multiple validities legitimation,27 validation requirements must be met for all aspects of the study. According to Bryman,28 six specific purposes exist for mixed methods research: credibility (to enhance the integrity of the findings), context (using qualitative research providing
contextual understanding coupled with generalizability, external validation, or broad relationships among variables), illustration (using qualitative data to illustrate quantitative findings), utility (to improve the usefulness of findings), confirm and discover (to generate hypotheses using qualitative research that is then tested with quantitative research), and diversity of views (combining researchers’ and participants’ perspectives). With true mixed methods, the study must have at least one point of integration. This integration occurs by merging the two data sets, connecting analyses of one set of data to the other, embedding one form of data within the larger design, and using a framework to bind the data sets. Refer to Table 3 for the four types of mixed methods designs.

The differences between quantitative and qualitative measures extend beyond the presence or absence of numbers. Quantitative research aims to answer questions of causality. In contrast, qualitative research focuses on answering the whys and hows. Although qualitative research often involves smaller samples sizes, the concept of saturation is used to determine data collection completeness. Data saturation is a measure of rigor that is used to estimate qualitative sample sizes and is a criterion for discontinuing data collection (ie, how many qualitative interviews must be conducted before no new data are found?).

**SUMMARY**

By selecting research methods aligned with Kirkpatrick’s levels of training evaluation, researchers can create stronger evidence of student learning when evaluating the effectiveness of teaching innovations. When designing or evaluating an educational research study, authors and reviewers should clearly define the learning outcome and ensure the tools and study outcomes match these learning objectives. Researchers should also move past gathering perception data alone and design rigorous methods to assess long-term retention of learning. They should consider mixed method approaches when appropriate and triangulate data from multiple sources when drawing conclusions. By carefully designing educational research projects, faculty can improve the quality of evidence for student learning.

**REFERENCES**


34. Denzin NK, Lincoln YS. Introduction: Entering the field of qualitative research; 1994:1-17.


### Appendix 1. Benefits and Limitations of Assessment Tools/Measures Categorized Along Kirkpatrick’s Four Levels of Evidence (Reaction, Learning, Behavior, Results)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Benefits</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reaction</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-Assessment</td>
<td>Student reflection required by ACPE standards; can be triangulated with other measures</td>
<td>May not include feedback to students, which could result in students’ viewing as busy work; if the students’ self-assessments are evaluated, it may require a large number of faculty to grade them; needs standardized rubric to ensure consistency among multiple graders or multiple students self-assessing; need for training of graders for consistent use of rubric; students may be weak at self-assessing</td>
</tr>
<tr>
<td>Perceptions, Confidence</td>
<td>Can be measured by pre/post and follow-up survey questions to track changes over time</td>
<td>Perceptions or confidence as measures alone do not ensure attainment of knowledge or skills; students may be weak at self-assessing; perceptions do not signify that learning occurred</td>
</tr>
<tr>
<td>Satisfaction</td>
<td>Can be useful to evaluate using a survey since low satisfaction may create a barrier to implementing a novel approach and high satisfaction may increase acceptance of new approach and facilitate implementation</td>
<td>Can be confounded by faculty personality or grading outcomes; students can be biased against new or novel approaches to teaching or assessment; could be confounded by elective or required course; does not imply learning occurred</td>
</tr>
<tr>
<td>Course Evaluations</td>
<td>Uses standardized questions that can be tracked over time; additional questions could be added to address study objectives; could include asking open-ended questions that would allow for qualitative thematic analysis</td>
<td>Can be impacted by low response rates, which could affect generalizability across all students; can suffer from gender bias; students responses can be influenced by timing of collection (eg, if given at completion of final exam, results could be focused on final exam vs. course); could be confounded by rewards for completion which could positively inflate results</td>
</tr>
<tr>
<td><strong>Learning</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quiz, Planned</td>
<td>Can be administered pre/post and at follow-up to track changes over time; quiz questions can be aligned with research objectives</td>
<td>Outcome may be confounded by students studying or preparing for quiz; high quiz grade does not imply long term retention, transfer or application; may impact long-term retention because of the testing effect</td>
</tr>
<tr>
<td>Quiz, Surprise</td>
<td>Can be administered pre/post and at follow-up to track changes over time; quiz questions can be aligned with research objectives; element of surprise may control for studying/preparing for the quiz</td>
<td>When conducted right after exposure, long-term retention, transfer, or application is not ensured; high quiz grade does not imply long term retention, transfer or application; may impact long-term retention because of the testing effect</td>
</tr>
</tbody>
</table>

(Continued)
Appendix 1. (Continued)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Benefits</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examination</td>
<td>Can be measured pre/post and at follow-up to track changes over time; exam questions can be aligned with research objectives; includes more content and questions than a quiz which may give more insight into knowledge</td>
<td>If longitudinally tracking a course over multiple years, course changes such as new faculty, new content, new questions, updated learning objectives may make it difficult to compare across years; high exam grade does not imply long term retention, transfer or application; when used as only outcome, conclusions are limited because students should gain knowledge as a result of being in the course</td>
</tr>
<tr>
<td>Cases(^{42})</td>
<td>Case questions with either open-ended or multiple-choice question prompts allow for assessment of higher-level knowledge application</td>
<td>Open-ended case question prompts may not be clear to the learner, resulting in incorrect answers due to clarity not lack of knowledge; case question writers may be overconfident that their question(s) were written at a high level; case specificity effect can occur, implying that success on any given case is specific to that case</td>
</tr>
<tr>
<td>TBL(^{43,44})</td>
<td>IRAT/TRATs allow for low-level knowledge assessment and the additional use of case questions with multiple-choice question prompts allows for assessment of higher-level knowledge application and ability to use evidence to justify solutions</td>
<td>Case questions are answered as a team, which prevent tracking of individual performance data; case question writers may be overconfident that their question(s) were written at a high level; case specificity effect can occur, implying that success on any given case is specific to that case</td>
</tr>
<tr>
<td>Course Grades(^{45})</td>
<td>Can track longitudinal changes over multiple cohorts/years of students; course objectives can be aligned with program learning outcomes</td>
<td>Can be confounded by participation grades or grade inflation; comparing course grades over multiple years may be difficult due to changes to grade components, course requirements, points distribution across years, and/or different course coordinators; high course grade does not imply long term retention, transfer, or application</td>
</tr>
<tr>
<td>PCOA(^{46,47})</td>
<td>Can benchmark across years in a program and across institutions; PCOA score report gives detailed feedback by area and topic; does not require study time for students</td>
<td>Limited number of questions per topic may result in a “case specificity” effect where students may have content knowledge but get the related PCOA questions wrong; high PCOA score does not imply long term retention depending on when administered, transfer or application of knowledge; PCOA measures knowledge not skills; PCOA blueprint may not be aligned with the study’s research objectives; poor performance may be confounded by exam stress, test length, unfamiliar testing software, etc.; could be influenced by testing practices of the institution (“prep” exam, incentives, etc.)</td>
</tr>
<tr>
<td>Behavior</td>
<td>OSCE (performance based)(^{48}) Authentic assessment method to assess both behaviors and knowledge; objective grading with standardized rubrics; confidence translatable to practice</td>
<td>May be a difficult assessment method to use because of extensive training required of case authors, standardized participants, and assessors; which can be costly; could be confounded by exam logistics, student stress, and standardized participants</td>
</tr>
</tbody>
</table>

(Continued)
<table>
<thead>
<tr>
<th>Measure</th>
<th>Benefits</th>
<th>Limitations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiential Learning (IPPE/APPE)</td>
<td>Authentic assessment method to assess both behaviors and knowledge; indicates if knowledge is transferable to practice</td>
<td>Interrater reliability can be difficult to achieve; generalized rubrics don’t allow for specific feedback; diversity of experiences make data analysis challenging; lack of grade variation or possible grade inflation</td>
</tr>
<tr>
<td>Results</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entrustable Professional Activities(^50)</td>
<td>Standardized language to define actions; can align study objectives to these measures</td>
<td>Not consistently used across programs; generalized rubrics don’t allow for specific feedback</td>
</tr>
<tr>
<td>NAPLEX/MPJE(^51)</td>
<td>Required of all graduates for licensure; developed by experts to assess knowledge of practice-ready examinees</td>
<td>Limited feedback provided to colleges/schools regarding graduate performance; limited feedback provided to examinees</td>
</tr>
<tr>
<td>Performance in practice(^52,53)</td>
<td>Ultimate outcome of training; can assess impact on patient outcomes</td>
<td>Limited avenue for data collection; no standard tool for measurement of success in various practice settings</td>
</tr>
</tbody>
</table>