AMERICAN PSYCHOLOGICAL ASSOCIATION

Assessment Guide for Psychology Teachers

2018

PREPARED BY

Working Group on Assessing Student Knowledge and Skills in Psychology

Dana S. Dunn

Alan J. Feldman

Stephanie A. Franks

Sayra C. González

Regan A. R. Gurung (Co-chair)

Jane S. Halonen

Miriya Julian

Rob McEntarffer (Co-chair)

Maria C. Vita

ASSESSMENT GUIDE FOR PSYCHOLOGY TEACHERS

This guide was developed by the Working Group on Assessing Student Knowledge and Skills in Psychology from the APA Summit on High School Psychology Education (July 2017). This guide was specifically written for high school psychology teachers, but its content may be useful to all psychology teachers (K-12, undergraduate, and graduate) as well as to teachers from other disciplines.

MEMBERS

Regan A. R. Gurung (Co-chair)
University of Wisconsin–Green Bay

Rob McEntarffer (Co-chair)
Lincoln Public Schools, NE

Dana S. Dunn
Moravian College, PA

Alan J. Feldman
Glen Rock High School, NJ

Stephanie A. Franks
Springboro Community City Schools, OH

Sayra C. González
Saint John's School, PR

Jane S. Halonen
University of West Florida

Miriya Julian
Clark County School District, NV

Maria C. Vita
Penn Manor High School, PA

We thank Suzanne Baker, Eric Landrum, Maureen McCarthy, Jennifer Schlicht, and Kristin Whitlock for their reviews of this guide.

CONTENTS

INTRODUCTION AND GOALS

How do you know students are learning what you want them to learn? How can you determine when a teaching strategy is working to produce the best effects or when it is time to make some changes to produce better results? How can you document your professional growth as a teacher? Assessment is the process of gathering evidence of student success and teaching effectiveness. Assessment outcomes are how we operationally define learning in our classroom—we attempt to measure what students know and are able to do through our classroom assessment processes.

The *Assessment Guide for Psychology Teachers* broadly conceives assessment to include both traditional strategies for testing and grading as well as approaches that focus more specifically on skill development through performance or "authentic" assessment. Sometimes the term authentic assessment is used to communicate the "real life" aspect of a simulation or other more complex performance expectations.

Teachers know that assessment is a vital part of the teaching and learning process. The overarching goal of this guide is to help teachers continue to build confidence in their assessment skills in order to provide useful feedback to students about their intellectual progress. These skills include creating or refining assessment strategies to evaluate students' knowledge and skills. This guide is also intended to help teachers assess the effectiveness of their teaching strategies in the spirit of continuous improvement. In the following sections, we provide guidance on assessment to facilitate useful, fair, and efficient assessment processes to benefit students and teachers.

It is our hope that this resource will not only provide much needed assistance in mastering evidence-based processes in teaching and learning but will also help engender both respect and enthusiasm for the power of assessment.

1. IDENTIFY THE PURPOSE OF ASSESSMENT

Carefully defining the purpose of an assessment clarifies what kind of assessment a teacher might deploy. For example, if the purpose of an assessment is a quick check for understanding or to use the data to make a teaching decision or get students to fix misconceptions, the assessment is going to look different from one designed to provide data for an evaluation at the "end" of learning.

Assessment experts often make a distinction between "formative" assessment, which represents more informal, low-stakes measurement or practice opportunities, and "summative" assessment, in which measurement directly contributes to a formal final

judgment of what has been achieved (e.g., a grade). Summative assessments tend to have higher stakes for the person being assessed. The distinction between formative and summative assessment applies to the performance of both students and teachers.

STUDENT PERFORMANCE

Formative assessments require students to demonstrate what they know in a low-stakes context, such as a practice exam or an ungraded pop quiz. Summative assessments for students define an outcome with higher stakes, such as a final grade or judgment about the success of a specific performance that will substantially influence a final judgment (e.g., a grade on a heavily-weighted project).

The following chart provides examples of student assessments for diverse purposes. The list is not meant to be exhaustive but may help teachers think broadly about assessment and making assessment choices.

Student Assessment

Formative assessment helps students develop their knowledge and skills in a low-stakes context.	**Examples** Pretests Practice examinations Project progress reports Self-assessment Peer feedback Ungraded checks for understanding Feedback on quality of ungraded class discussion
Summative assessment produces a high-stakes judgment about student knowledge and skills at the conclusion of a process.	**Examples** Major project completion Posttests Content assessments (e.g., unit tests, midterms, finals) Portfolio presentations

See Karbach (2014) for further clarification of differences and examples.

TEACHER PERFORMANCE

Teacher formative assessment provides feedback on how well teachers may have taught a concept or theory to help them determine which concepts or skills might need more work. In such low-stakes cases, the teacher may not provide individual feedback to students on how they did, since the question focuses on the "aggregate" performance—that is, how the group did as a whole. In contrast, teacher summative assessment produces a high-stakes outcome (e.g., a teacher might examine and report the patterns of student achievement for the year to justify claims of successful teaching to an administrator). Implications of teacher summative assessment are addressed toward the end of this guide.

Teacher Assessment

Formative assessment involves low-stakes activities that shape teaching direction drawn from student performance or stakeholder opinion	**Examples** Ungraded checks for understanding Patterns emerging in student discussion "Exit ticket": forming or answering a question that yields conclusions about mastery "One-minute paper": summarizing and synthesizing key or difficult ideas in a time-limited fashion through end-of-class free writing Midterm student evaluations of teaching Peer review Personal reflections
Summative assessment produces high-stakes outcomes regarding teaching achievement	**Examples** Grade distributions Patterns of student outcome achievement Pretest/posttest comparisons End-of-term teaching evaluations Teaching award competition processes

The categories are not mutually exclusive. For example, a pretest to establish a baseline can inform students about the limited scope of their knowledge before learning some targeted material while also identifying areas that may need attention as a teacher designs the learning experiences in the course. Similarly, a peer review conducted to help teachers improve (formative) might also serve as evidence of quality in a high-stakes decision, such as a teaching award competition (summative).

In conclusion, you need to know WHY you are assessing to make the best judgments about HOW to gather evidence of student achievement. The next section explores how to choose the right tool for the right assessment question.

2. CHOOSE AN APPROPRIATE ASSESSMENT FORMAT

After defining the purpose of the assessment, the next step is choosing an appropriate assessment format for your items/tasks. Assessment specialists refer to this step as the "match" between assessment format and measurement goal. An appropriate choice takes into account the following dimensions:

DIRECT VERSUS INDIRECT METHODS

Direct methods measure some observable aspect of performance, whereas indirect methods assess an opinion, belief, or attitude. For example, applying a rubric to a project report represents direct measurement of a student performance; an indirect measurement would explore student satisfaction with the work. Generally, direct performance measures carry more weight as evidence of achievement than do indirect measures. However, some questions may be more suited to indirect than direct strategies (e.g., "Rank order the value of contributions of members of your project group").

OBJECTIVE VERSUS SUBJECTIVE STRATEGIES

Assessment experts distinguish between objective and subjective formats. Objective strategies constrain how students can show their understanding. For example, multiple-choice and fill-in-the-blank approaches limit the options students can choose to reflect their knowledge or acquired skills. Grading of objective work tends to be easy and typically is evaluated electronically; however, most instructors recognize that significant time must be invested in the design of a high-quality objective test and may focus on simpler cognitive skills, such as recall and recognition. Performance feedback in objective strategies tends to be expressed in "quantitative" terms, such as total of correct responses or percentiles.

In contrast, subjective strategies (e.g., open-ended essay questions) are typically categorized as more "qualitative"; these tend to be easier to create and harder to grade, since a grader must render individual attention to produce a fair conclusion about what the students have mastered in content or skills.

THINKING SKILL LEVEL

Traditional testing and performance assessment tend to focus on different levels of cognitive complexity. Traditional testing measures currently emphasize the most basic of cognitive abilities. High school teachers report that their school systems often endorse a formal framework to promote cognitive development. Two popular frameworks in current high school practice are **Webb's depth of knowledge model** and **Bloom's taxonomy.**

Webb's depth of knowledge

Webb's depth of knowledge (DOK) approach uses four levels of analysis to cast cognitive levels as going from shallow to deep abilities (Webb, 1997). The first level provides an overall description of cognitive levels in sequential development: Students acquire knowledge, stressing more basic skills of recall and memorization; students use knowledge, emphasizing skills such as explaining, analyzing, or evaluating; and students extend their thinking by solving complex problems, completing an investigation, or making creative connections. The second-level analysis provides more detail by taking into account transitions from one major function to the next. These terms may be more familiarly used in school settings that support the DOK as their primary cognitive framework.

The levels are as follows:

- Level 1: *Recall/recognition* emphasizes demonstrating acquired knowledge by retrieving relevant information or procedure.
- Level 2: *Skill/concept* use involves acquiring and using information or conceptual knowledge.
- Level 3: *Strategic thinking/reasoning* incorporates greater complexity that uses reasoning, planning, or coping with more than one possible answer.
- Level 4: *Extended thinking* is the deepest level, which entails the completion of an investigation or project exploring complicated problems.

The final DOK level of analysis focuses on action verbs that might be deployed in developing student learning outcomes, such as analyze, generalize, and solve (see Figure 1 for a depiction of the overall skills, skill levels, and specific cognitive abilities).

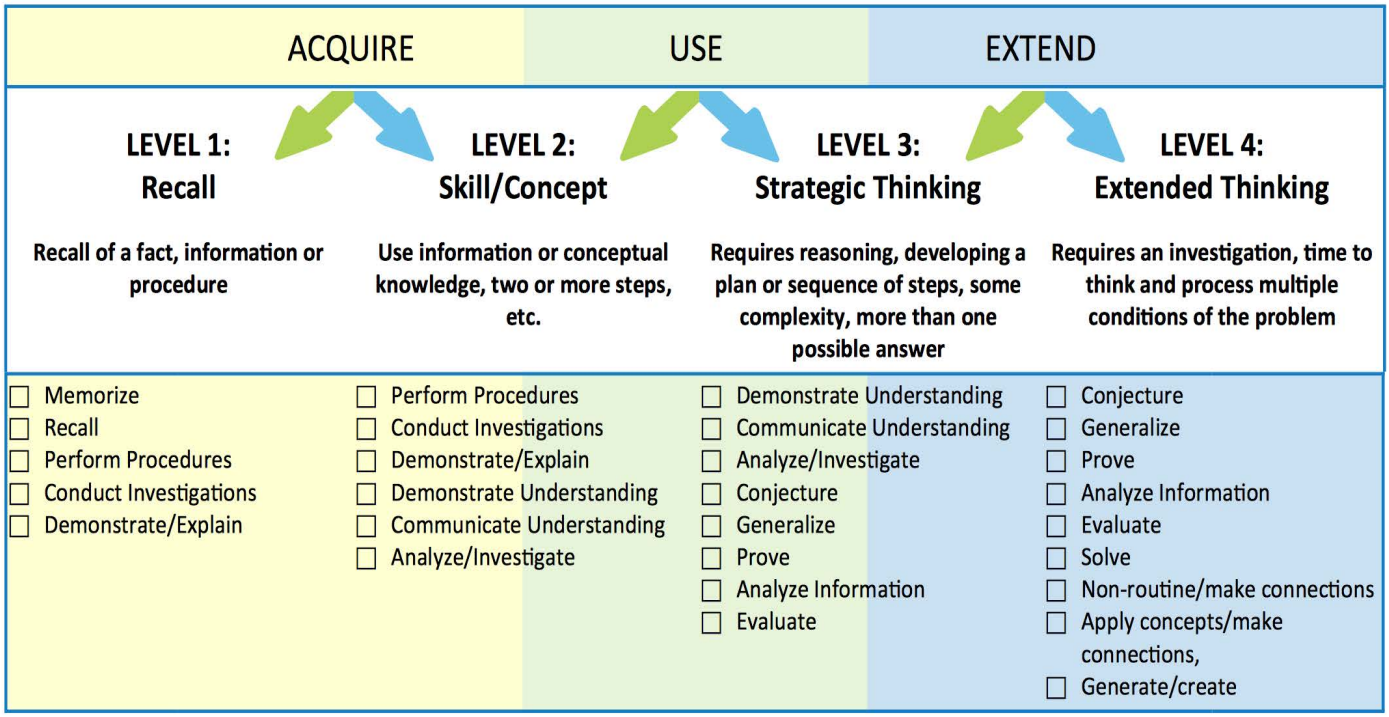


Figure 1. DOK levels, transition characteristics, and cognitions.
From *Webb Leveling: Expectations for Student Performance (n.d).*

Bloom's Taxonomy

Another popular cognitive framework, Bloom's Taxonomy (Bloom, 1956; revised by Anderson & Krathwohl, 2001), provides a six-level framework that describes "lower order thinking" through "higher order thinking" (see Figure 2). Like the DOK, Bloom's Taxonomy emphasizes basic skills as lower order and highlights creative thinking as the higher order skill. Bloom's Taxonomy tends to be used more often in college settings.

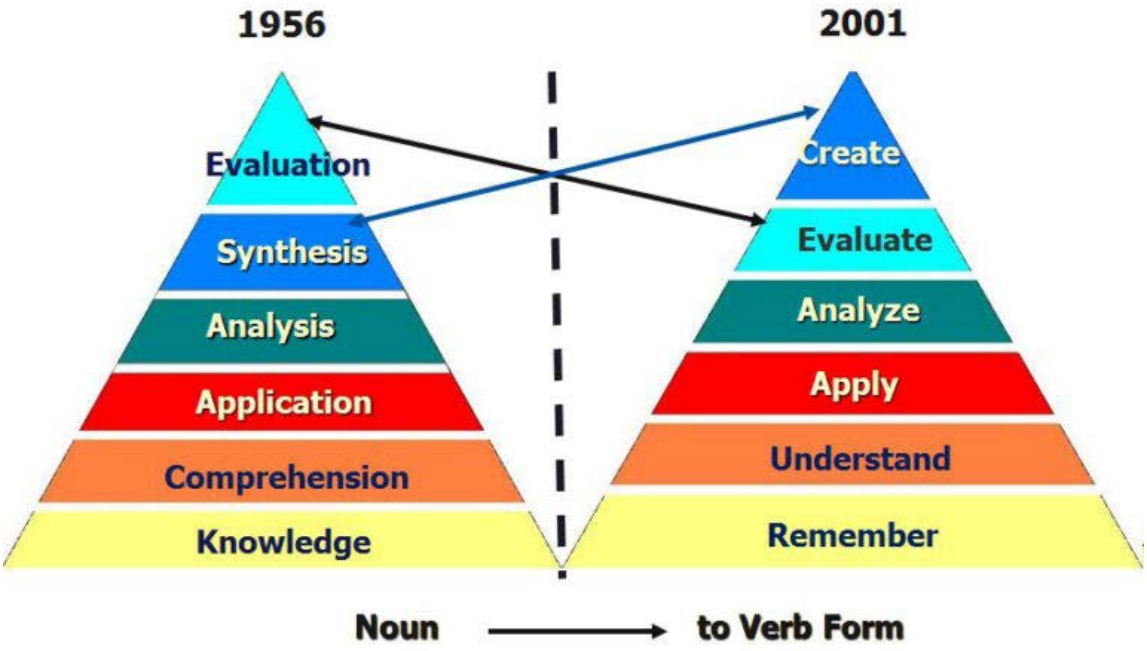


Figure 2. Bloom's Taxonomy.
From Wilson (2016).

Teachers should keep cognitive complexity in mind when designing both pedagogical activities and assessment. For example, “forced-response” items (like multiple-choice and fill-in-the-blank questions) lend themselves to assessing lower level cognition, such as recalling facts and understanding and applying concepts. Skilled test creators can develop multiple-choice items that tap more complex thinking skills, such as analyzing and evaluating, but these are much harder to create. Essay questions more typically ramp up cognitive demand to the mid-range skills of applying knowledge, analyzing information, and making evaluative judgments. These approaches contrast with highest level cognitive demand found in strategies that focus on project development in which students typically create some kind of product or performance. Performance assessment or authentic assessment are terms used to address measuring the outcomes related to higher level cognitive demands.

Beyond the simple characterization of shallow vs. deep (DOK) and lower vs. higher order (Bloom), there is significant overlap in the approaches of these two frameworks. Note how both approaches foster an understanding that students’ cognitive performances can range from relatively simple, low-demand thinking to that which is much more complex. Both approaches highlight how the nature of evaluating student learning depends on the cognitive outcome targeted by the teaching strategy (see Table 1).

Depth of Knowledge	Bloom’s Taxonomy	Assessment Focus
Acquire	Remember Understand	Testing
Use	Apply Analyze	Performance Assessment
Extend	Evaluate Create	Performance Assessment

Table 1
Comparing the Goals and Corresponding Assessment Strategies of Webb’s Depth of Knowledge to Bloom’s Taxonomy
Adapted from Brookhart (2017).

SOFTWARE VERSUS HUMANWARE

The degree to which evaluation might be automated also influences the choice of an assessment strategy. Objective measures easily lend themselves to grading that does not require ongoing human attention once the programming bugs have been resolved. Progress has also been made in applying algorithms to produce automated feedback on such measures as essay writing. Substantial work will be required to create the evaluative structure, but a "plug and play" assessment strategy has a lot of appeal for time-strapped teachers. Some systems may even include automated notification of results to students, including relative performance standing in the class, to reduce teachers' workload further. We refer to the work of the technology group from the APA Summit on High School Psychology Education for current examples of technological supports for pedagogy and assessment (e.g., White, 2018).

ACCOMMODATIONS

Accommodations are changes to an assessment (or the way an assessment is administered) that help students better represent what they know or can do. Accommodations are different from modifications:

- **Modifications** change what is being measured (e.g., changing essay questions to fill-in-the-blank questions for students with limited writing ability).
- **Accommodations** change how learning is measured, not what is being measured (e.g., allowing students extended time to finish a test).

Some students have documentation (individual education plans, etc.) that describe what accommodations are recommended (or even required) during testing. These plans are often associated with diagnosed learning disabilities or other learning issues. Usually these accommodations can be provided in ways that do not disrupt or inconvenience the rest of the class. Some accommodations may require testing the student in a small group or an individual setting. Ideally, teachers and students will discuss testing at the beginning of a course and decide how best to provide the accommodations to which students are accustomed and which allow them to demonstrate what they have learned. Students usually have had experiences using these accommodations in other classes and should be able to help plan how accommodations can be used efficiently during testing.

Different schools or institutions use different lists and labels for accommodations, and it would be impractical to try to include a complete list here. Most accommodations can be organized into one of the following four categories:

- **Presentation:** A change to the way test items are presented.
  Example: Students might listen to an audiofile rather than read a text passage before answering questions about the passage.
- **Response:** A change to the way a student responds to test items.
  Example: Allowing a student to submit a typed response rather than a handwritten one.
- **Setting:** A change to the environment in which a student completes an assessment.
  Example: A student tests in a separate room with fewer distractions, or in a smaller group.
- **Timing and scheduling:** A change to the amount of time a student has to complete a task or a change to the schedule.
  Example: Providing extra time on tests or allowing a student to take breaks.

3. ADOPT/WRITE/REVISE ASSESSMENT ITEMS OR TASKS

You know about what you want your assessment to do (purpose) and how you want to do it (assessment format); now is the time to adopt, revise, or create your items/tasks. In this section, we provide advice on selecting, revising, and designing high-quality assessment strategies.

HOW DO I DEFINE SUCCESS IN STUDENT PERFORMANCE?

Any assessment is likely to produce variable student performance. How you establish your own expectations for success will shape the nature of the feedback you provide to the student.

Norm-referenced grading

Traditional grading practices typically assume that collective student effort will be distributed normally. The industry standard for grading assumes a bell curve despite the fact that much has been written about grade inflation over recent years. Generally speaking, faculty derive some comfort from the standard practice of assigning grades based on accuracy percentages (e.g., an "A" is 90% or above of available points). Assuming the existence of a competitively derived bell curve tends to support an expectation that some students will do very well and some will do poorly (see Figure 3).

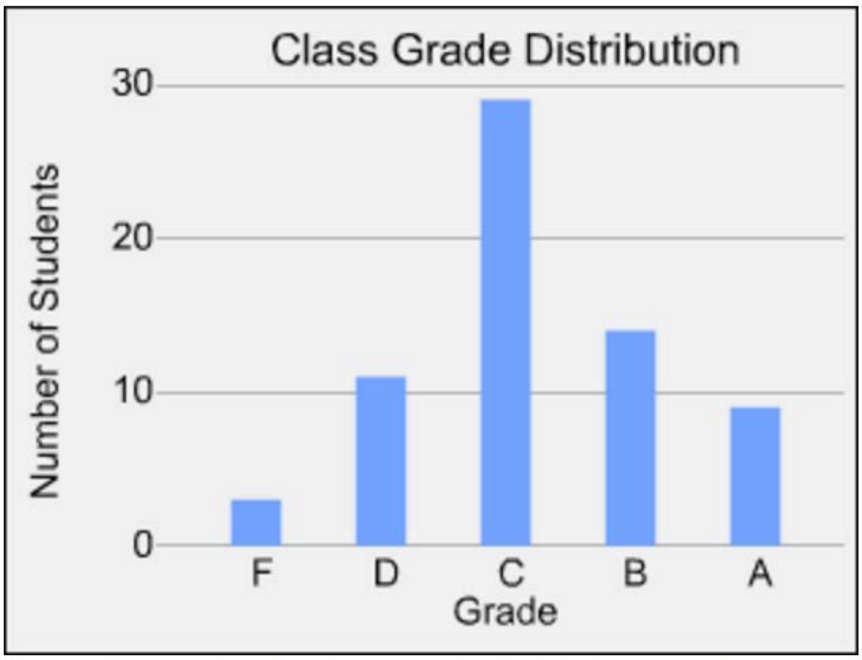


Figure 3. Example of a class grade distribution.
From Gramoll, K. (n.d.)

Criterion-referenced grading

An alternative to norm-referenced grading, criterion-referenced approaches tend to be used in situations involving performance. In this approach, the teacher determines what appropriate performance should look like ("student learning outcomes") and then describes that performance using action verbs to constitute criteria. For example, in exploring psychology's use of the scientific method, a good exemplar of a measurable outcome is "Interpret simple graphs and statistical findings." Assessment experts refer to this strategy as criterion based because it describes the essential elements of performance. When students render a performance, each criterion is judged on a continuum of high to low quality. To support instructional clarity and consistent grading, the assessor develops a "rubric," a matrix that links the list of criteria to expected levels of performance. A simple performance rubric might designate three performance levels:

- Exceeds expectations (alternative terminology: Exemplary, Outstanding, Mastery, Excellent)
- Meets expectations (Proficient, Competent, Satisfactory, Good, Acceptable)
- Does not meet expectations (Insufficient, Novice, Developing, Emerging, Needs Work)

Here is how this scheme might apply to the example of a specific criterion about effective writing mechanics in a paper:

	Does not meet expectations	**Meets expectations**	**Exceeds expectations**
Writing mechanics	Commits many errors in grammar/ spelling	Commits a few errors in grammar/ spelling	Commits no errors in grammar/spelling

Performance-based assessment does not presuppose that a certain percentage of students will fail or necessarily promote competition for a grade but typically measures more complex levels of cognition. For these reasons, many educators strongly prefer criterion-based approaches over norm-referenced strategies. In addition, criterion-referenced grading tends to move the emphasis toward the development of skills in the context of the discipline's content, which is likely to produce longer lasting gains from the learning experience than do norm-based, content-focused approaches.

Although it is useful to set some performance benchmarks ahead of time to determine whether student learning has achieved the objective you had in mind, seasoned assessors recognize that criterion-referenced grading tends to produce continuous refinement of the process. Student performance provides clues as to whether the instructions were clear or the expectations too high. We address the specifics of rubric design later in this section.

HOW DO I CHOOSE HIGH-QUALITY STUDENT LEARNING OUTCOMES?

You do not have to start from scratch—APA has done substantial work on generating relevant student learning outcomes and supporting pedagogy for introductory courses. The prevailing standards are represented in the *National Standards for High School Psychology Curricula* (APA, 2011). The curriculum design detailed in this document provided the infrastructure for the exemplars of good assessment design that were developed at the APA Summit on High School Psychology Education. APA requires that official policy undergo evaluation and, if appropriate, revision every 10 years. The National Standards will expire 2021; the next revision is expected by August 2021. In 2017, an APA task force specifically addressed the outcomes of the introductory psychology course in *Assessment of the Outcomes of the Introductory Course in Psychology* (APA, 2017). Although this work was originally geared to address college-level courses, the curricular architecture offered in this report may influence the direction taken in the next revision of the *National Standards*.

Here are some tips for crafting your own student learning outcomes specific to the projects you create and subsequently assess:

- Use language crafted for ease of understanding for the individual whose performance will be measured.
- Limit the number of criteria on which your evaluation will focus to avoid overwhelming the student.
- Limit the number of levels to distinguish quality of performance so that you won't be trapped into making distinctions that are too hard to draw.
- Focus on action verbs, such as analyze or evaluate, and avoid verbs for which it is harder to provide evidence, such as understand or appreciate.
- Observe local practices in terminology. For example, some schools discourage the use of the word "fail" to describe student performance and may advocate for the less harsh "developing" to communicate that the student needs more development.
- Think about student learning outcomes as "evolving" because student performance helps establish gaps or problems with the protocol.

An example of a student learning outcome taken from the *National Standards* (APA, 2011) is: Students will be able to identify factors that influence encoding. The 2017 *Assessment of the Outcomes of the Introductory Course in Psychology* mentioned previously provides numerous student learning outcomes for teachers.

HOW DO I WRITE HIGH-QUALITY MULTIPLE-CHOICE QUESTIONS?

Some general tips about constructing multiple-choice questions that produce the best measurement of learning include the following:

- Focus on the disciplinary content you covered, or content validity will suffer (see Section 4 for a discussion of content validity).
- Select test questions that reflect important concepts rather than trivial details.
- Include test items that might be a bit easier early in the test to build student confidence.
- Incorporate more difficult/challenging items to help differentiate students who learned the material more effectively.
- Write items that depend on course experiences and/or course readings (e.g., can someone who has not taken your course answer the question?)
- Choose the number of items according to how many slower readers can manage in the allotted time.
- Craft well-worded and easy-to-understand items (i.e., avoid the use of jargon).
- Be careful with items that require a cultural context (e.g., not everyone may understand multiple-choice questions that involve the Super Bowl or World Series).
- Avoid having the longest answer or the shortest answer as the correct answer.
- Derive distractors (wrong options) from related content.
- Include no more than three or four distractors (incorrect answers).
- Use humorous alternatives carefully as these may harm performance for some distractible students.
- Consider rewriting “Which of the following is not” stems to emphasize the content rather than the logical relationships between alternatives.
- Avoid using “all of the above” and “none of the above” as options, since students are technically right if they choose only A or B but will be penalized for lack of comprehensiveness or failure to read thoroughly.
- Consider using a “set question” in which a more complex stem (question), short writing passage, or description of research, etc., serves as the stimulus for multiple questions.

Additional information along with sample questions that illustrate multiple-choice questions designed to address different levels of cognitive complexity can be found in Appendix A. This document represents advice for teachers who develop items for the Advanced Placement Psychology exam from the Educational Testing Service, although the advice is relevant to any psychology teacher.

HOW CAN I DEVELOP ESSAY QUESTIONS THAT ENCOURAGE THE RIGHT COGNITIVE COMPLEXITY?

Essay questions can facilitate students' development of greater cognitive complexity if the questions are well designed. Consider the following tips for optimizing the results of essay question strategies:

- Write the stem as clearly and concisely as possible.
- Select the most appropriate verb that corresponds to the cognitive level you want to see. For example, list and describe will produce less complex answers than explain or compare and contrast.
- State explicitly whether students can use notes to support their thinking and writing.
- Develop an objective, measurable scoring guide before you begin grading. Be clear about what counts beyond conceptual clarity and accuracy (e.g., good grammar, complete sentences). In some cases, you may able to share the guide or rubric with students ahead of time to facilitate their best answers.
- Grade essays blind (cover up student identification indicators) to protect against "halo effects" (bias about specific students that might influence teacher evaluation).
- Be direct, but gentle, with feedback. Be sure your comments address what the student did right rather than only what the student did wrong.
- Emphasize the biggest or most important improvements that need to take place in developmental feedback so students will not feel overwhelmed. Grading a response to promote student growth is different from editing to perfect the production.
- Ask for a follow-up reflection on what students learned or accept a revision for partial credit to improve the likelihood that they will read and benefit from feedback.
- Consider the initial use of an essay question as field testing. The results will allow you to revise the question and the rubric in relation to exposed weaknesses in the future.

For further information on writing effective essay questions, see Reiner (2002).

WHEN SHOULD I USE TRUE–FALSE QUESTIONS?

True–false testing enables broad and fast coverage of targeted content. However, from an assessment standpoint, true–false testing may not be very informative about what students actually have learned, since each question offers a 50% chance of scoring accurately without really knowing the answer.

In addition, psychology does not always lend itself readily to the black-and-white world of true–false questions. The evolving nature of science leaves open the possibility that what is taught as truth at one point may need to be revised as we learn more about phenomenon. Here are some tips about preferred practices when using true–false questions:

- Use true–false items when you need quick feedback on basic, relatively simple concepts.
- True–false items can expose students' misconceptions about erroneous ideas they hold about behavior, as demonstrated in empirical literature.
- You may be able to use true–false items with complex concepts to illustrate concept complexity when a simple answer is not appropriate.
- Encourage students to justify their answers to push more complex thinking (e.g., reasoning).

WHAT ARE THE BEST STRATEGIES FOR PERFORMANCE RUBRIC DESIGN?

In performance assessment, teachers have the opportunity to evaluate both content and skills. Some assessment experts describe the process as evaluating a performance in the context of targeted content. Rubrics differ in their complexity and use to achieve desired student learning outcomes. Some relevant distinctions follow.

Holistic vs. analytic design
Holistic rubrics focus on the quality of the entire performance. Because they tend to be less detailed, these can be better deployed when the expected performance is simple or the students are less able to handle complexity. Grading transpires when the assessor looks at the pattern of performance and decides how that pattern might justify an overall grade. In contrast, analytic rubrics produce a judgment about the work quality by looking at its component parts (e.g., criteria). For grading in the analytic rubric, each criterion contributes an explicit value toward the overall grade. The more specific the rubric, the greater likelihood that judgments derived using the rubric will be reliable.

Single use vs. generic use

Teachers will create some rubrics in response to a given set of instructions about student performance and may use the rubric just once in relation to this circumstance. Some rubrics, however, may have more generic applications. For example, a well-designed writing quality rubric might be used in relation to multiple writing projects during a student's development.

Single opportunity vs. multiple opportunity

A rubric can be used to promote an appropriate revision approach in projects for which revision is needed. The initial application of the rubric results in feedback that should be addressed in any subsequent submission.

Single-point rubrics

An emerging popular design is an analytic rubric that restricts the evaluative dimension of the rubric to just three levels (see Figure 4). The target criterion or standard for the performance occupies the middle column. Achievement at an acceptable level can be circled. If unsuccessful, the evaluator can make notes for improvement in the "Concerns" column. If particularly well executed, that achievement can be noted in the "Advanced" column.

Concerns *Areas that Need Work*	**Criteria** *Standards for This Performance*	**Advanced** *Evidence of Exceeding Standards*
	Criteria #1: Description reflecting achievement of mastery level of performance	
	Criteria #2: Description reflecting achievement of mastery level of performance	
	Criteria #3: Description reflecting achievement of mastery level of performance	
	Criteria #4: Description reflecting achievement of mastery level of performance	

Figure 4. An analytic rubric in which the evaluative dimension is restricted to three levels.

From Gonzalez (2014).

ADDITIONAL TIPS FOR CREATING POWERFUL AND EFFECTIVE RUBRICS

- Avoid making a dense (too many words in each cell) or too complex (too many elements) rubric. Try to invoke “Miller’s magic number” (7 + / - 2) to limit how many criteria or judgment levels you use.
- Use language that speaks to the student. Avoid the use of terms they will not understand.
- Consider getting feedback from students on how effectively the rubric communicates your expectations either before you finalize the rubric or after its first application.
- Focus on providing feedback that students can use to improve. Overwhelmingly critical feedback will be dismissed or ignored.

HOW CAN I CREATE MEANINGFUL INDIRECT ASSESSMENTS?

Indirect assessments generally concentrate on perception or opinion rather than on skill development. In an era in which nearly all businesses follow up transactions with customer satisfaction surveys, we can assume that most students will understand the value of asking their opinions about teaching and learning processes. Some indirect assessments fare better using open-ended items, although the data may be more difficult to aggregate at the conclusion of the questions. Consequently, many indirect assessments tend to be more quantitative, using a Likert scale ranging from 1 (strongly agree) to 5 (strongly disagree), for example. Best practices for creating opinion surveys include the following:

- Avoid using biased or loaded questions.
- Consider what type of question best suits your purpose.
- Limit rating scales to no more than seven points if using a Likert-type scale. Length and ease of completion will influence response quality.
- Temper conclusions according to the response rate. For example, if fewer than 25% of individuals targeted offer feedback, the information gathered may be of questionable validity.

HOW CAN I INCORPORATE MEANINGFUL SELF-ASSESSMENT?

Teachers often praise the use of self-assessment (e.g., having students submit a judgment about the quality of the work they have accomplished) as part of the assessment process in promoting the best learning outcomes. Using this approach may facilitate students’ taking greater responsibility for the quality of their work and improve internal locus of control.

The simplest way to promote self-assessment is to ask students to identify whether they have achieved their best work. If time allows, some teachers might reward an honest but negative self-assessment (e.g., “This is not my best work”) with additional time in order to drive home the importance of students’ taking pride in their work.

A slightly more sophisticated approach is to ask students to explain a strength and/or weakness in the work. Some teachers have found that sharing the evaluation rubric with the students and asking them to fill the rubric out according to what they think they have achieved is a good way to stimulate stronger performance.

For more information and classroom activities on incorporating metacognition into the psychology classroom, please see the resources developed by a working group at the APA Summit on High School Psychology Education that focused on metacognition skills. These resources will be available online by the fall of 2018 through the APA website.

4. EVALUATE THE EFFECTIVENESS OF ASSESSMENTS

Once you have designed and implemented your assessment, you need to establish whether the results have value: Are you measuring what you want to measure? Is the measurement stable? Psychometrics is the branch of psychology that focuses on the quality of test design and implementation. Specifically, a good assessment instrument must produce valid (truthful) distinctions as well as reliable (consistent) findings.

A good measuring strategy must be both reliable and valid. For example, if you weigh yourself and your scale is in good working order, you should be able to have confidence that the scale is producing accurate results; the measurement is both valid (it measures your actual weight) and reliable (it produces a consistent estimate every time you step onto the scale). Suppose your scale goes out of whack and adds 10 pounds to your real weight every time you step on the scale. In this case, your measurement is reliable (the scale adds 10 pounds every time you step on the scale), but it is not valid (the measurement is going to be off by 10 pounds every time). Poor tests produce problematic validity and reliability.

An important aspect of psychometrics also addresses whether the assessment produces fair judgments—that is, all students should have an equal opportunity to demonstrate their knowledge regardless of advantage or disadvantage. In this section, we explore strategies for producing the most satisfying results.

WHY YOU NEED TO UNDERSTAND CORRELATION COEFFICIENTS

Judgments involved in evaluating the quality of quantitative assessment involve comparing one set of numbers to another set to establish their strength of association (i.e., comparing data from a newly developed test to a different, well-established test to check validity). Correlation coefficients determine the strength of association between those two sets of numbers and are expressed numerically as a number between -1.0 and +1.0. The closer the correlation is to +1.0, the more confidence you can have that the two sets of data are strongly correlated, which would be the goal in deriving most correlation coefficients involved in psychometrics. Figure 5 illustrates how to interpret correlation coefficients.

Figure 5. Correlation coefficient showing strength and direction of correlation.

For example, if two data sets produce a correlation (strength of association) of +.89 and another comparison of two data sets produces a correlation of +.44, then you can say with confidence that the first comparison documents a stronger association between the compared data sets. As the correlation coefficient approaches zero, the association becomes weaker. Correlations that are on the negative side demonstrate that when the scores in one data set are high, the scores in the other data set are low. For example, a correlation between two data sets of -.78 would be a stronger association than a correlation between two other data sets of -.52.

In most cases, psychometricians strive to produce correlation coefficients that are on the high positive end of correlations. Although it is unlikely that correlations involving human behavior ever produce a perfect positive correlation of +1.0, the closer the correlation is to that number, the stronger the evidence is for determining test quality.

WHAT YOU NEED TO KNOW ABOUT RELIABILITY AND VALIDITY

You want to design effective assessment strategies so that you will be confident in the judgments you draw from the results. How do reliability and validity apply to assessment decisions?

Reliability
Several kinds of reliability determine whether you can have confidence in the consistency of an assessment strategy.

- **Test–retest reliability.** A test should produce similar results on subsequent use of the same test. For example, if you test a group of students on concepts on Wednesday, the test should produce the same results on Friday if the students have not been exposed to new information in the interval.
- **Parallel forms reliability.** You might turn to a large test bank of multiple-choice questions to produce a test. In a very large class it might be worth your trouble to construct two separate tests from the test bank so that students seated next to each other are discouraged from copying, since the neighboring student's test would be different. Running a correlation coefficient between the two data sets will tell you whether the tests were equally difficult and might dictate that some adjustment should be made to make their scores equivalent if the correlation is not strong (see earlier section on correlation coefficients).
- **Interrater reliability.** When a test has indisputable answers (a strong answer key), there is no need to determine whether different raters would come to different conclusions. However, when a test circumstance depends on expert judgment, such as grading an essay question, experts may differ in their judgments even when they have been trained on the rubric. A high positive correlation coefficient provides assurance that the experts are coming to the same conclusions about the quality of a performance.
- **Split-half reliability.** This strategy provides some assurance that the elements of a given test are consistent. Test items are divided in half, and the results of the first half are correlated with the second half. Internally consistent tests produce high positive correlations.

Validity

Several kinds of validity inform whether you can have confidence in the truthfulness of the measurement.

- **Face validity.** Does the test look like it is a plausible measure of what it claims to measure? Although this type of validity does not involve statistics, a good test will motivate test takers because they believe the content of the test appears to be a good match for the knowledge they are expected to demonstrate.
- **Content validity.** Do all of the elements of the tests correspond to the appropriate content used to assess learning? For example, if your test is supposed to cover chapters 5–7 and a few questions from chapter 8 end up on the exam, you have violated content validity.
- **Construct validity.** Does the test actually measure the psychological constructs you intended? For example, if you construct an essay test in which the vocabulary used in the questions far exceeds the knowledge base of students to be tested, then you will not be able to assess what students actually know about the targeted content. You will have violated construct validity.
- **Criterion-related validity.** To what degree do your test results correlate with some other relevant measure? For example, how well a student performs in the first semester of college should show a high correlation with college entrance tests if the tests are sound. Sometimes this form of validity is referred to as predictive validity because the results of a test can predict some external outcome (e.g., success on the job, likelihood of graduating).

HOW DO I ENSURE THAT MY MULTIPLE-CHOICE QUESTIONS ARE SOUND?

If you choose to assess using multiple-choice questions, you want to ensure that your test questions are effective in revealing what students know and how they think. You have several options for how you can evaluate whether your multiple-choice questions are doing the job:

Low-tech indicators

In this approach, you can determine by listening or observing where problem areas might lie.

- Students may ask clarification questions about an item during the test. The student's vantage point may alert you to an inadvertent error or poorly constructed item, or vocabulary or wording issues, that can be fixed before scoring or later when you want to revise the item.
- If the normal distribution of scores is badly skewed or lopsided, it means that the test may have been too rigorous or too easy.

- If all students get the item right, the item doesn't offer much discrimination among learners, so it may be of limited value for differentiating how well students know the material. However, an item or two may be purposefully included, especially early in the exam, to build confidence.
- If all students get an item wrong, either the question is poorly designed or the target content could have been presented better.
- If several students who mostly do well on the test as a whole get an item wrong, it is an indication that something could be amiss, but it doesn't necessarily guarantee that the item is bad.

High-tech indicators

If you have the capacity for item analysis in the automated grading system you use, you should be able to tap statistical data to help you determine where improvements might lie.

- **Item analysis.** This information displays the distribution of all student answers, which can inform you about which distractors seemed to be most appealing. Ideally, the majority of students should pick the correct answer.
- **Difficulty score.** This is simply the percentage (usually expressed as a decimal) of students who answer an item correctly. Test items with difficulty scores that are too low (indicating that very few students found the correct answer) or too high (indicating that almost all students got the item right) may need to be rewritten to better measure what you want to measure.
- **Discrimination index.** This statistic provides a calculation that compares students' overall scores to how they did on individual items. A test item with high discrimination "differentiates" between students well: Students who did well on that test item did well on the test overall, and students who did poorly on that item didn't do well on the test overall. If you find test items that have discrimination indexes close to zero or that are negative, you may want to rewrite those items because there may be something about the wording of the stem or the possible answers that confused your students.

Enrichment strategies

Some teachers find they can use student errors as the basis for pushing students' understanding in a second stage of review. For example, you might offer students an opportunity to earn back a portion of lost credit for an item by explaining why they made the wrong choice to begin with and why the correct answer would have been better (these are sometimes call "test correction" strategies).

HOW DO I KNOW IF AN OVERALL TEST MEETS MY NEEDS?

- **Get feedback from students.** They can pinpoint elements of the test that were confusing or unexpected to help you with revising the test for future use.
- **Look at the grade distribution.** Chances are good that a sound quantitative test will produce a distribution of scores. You may not be looking for an exact "bell curve" (see the How Do I Define Success in Student Performance section on page 13), but usually you would expect a valid test to produce a range of scores in your class. A very skewed (lopsided) distribution may demonstrate that the test is too hard or too easy.

WHAT ARE OTHER STRATEGIES YOU CAN USE TO IMPROVE TEST VALIDITY?

- Review how your test intentions fit with the goals and objectives you formulated for the course.
- Proofread your test carefully to catch typos, avoid queuing of correct answers, verify the test questions are numbered properly, and so on. Tip: Consider having students generate sample multiple-choice items, answer them together in class, and then use a sampling of the student-generated items on the actual test.
- Get students or some other external reader to look over your assessment stimulus to determine if instructions are clear (i.e., the wording is appropriate and easy to understand).
- If relevant, compare your results to other targets that correspond to your test plans.
- Consider examining item statistics (see Low- and High-Tech Indicators sections on pages 25-26) to identify items that need revision.
- Consider using a "think aloud" strategy: Ask students to choose the answer they think is correct, and provide them with space to explain why they chose that answer. You can use students' input to change items so that they better measure the kinds of thinking you want to assess.

HOW CAN I DETERMINE IF MY PERFORMANCE ASSESSMENT STRATEGY IS EFFECTIVE?

If your focus is performance assessment, there are many parameters to consider either in developing your own or in adapting someone else's assessment. The group of educators who met at the APA Summit on National Assessment in Psychology (2016) developed a rubric, included in Appendix B, to help you make good decisions about whether an instrument will meet your needs.

WHAT IF MY STUDENTS SEEM TO BE UNMOTIVATED TO PERFORM WELL ON THE ASSESSMENT?

A common problem reported in the assessment literature is that students may not be strongly motivated take assessments seriously. They may already have developed an attitude of learned helplessness that they cannot master what is required to do well in school, or they may not care much for the discipline they are studying. Some suggestions for ramping up motivation include the following:

- Ask students to self-assess.
- Involve students in the design of the assessment or evaluation rubric.
- Ensure that the assessment makes a reasonable contribution to the grade judgment.

HOW DO I JUDGE SUCCESS ON THE WHOLE?

When determining the overall success of an assessment or program, two strategies have emerged as prominent: the use of a target benchmark and achievement of an improvement rate:

- **Target benchmark strategy.** In this approach, you average the achievement of all students on the assessment measure. If the average exceeds 70%, you can feel reasonably confident that students are learning the knowledge and skills you have identified as important. Of course, bragging rights can attend to average performances that greatly exceed 70%. Adopting a 70% benchmark is a preferred solution for programs that are reasonably strong.
- **Improvement rate strategy.** Once a baseline level has been established, another strategy is to target a low-level percent improvement. For example, a program can declare a goal of 2% improvement in reaching a performance level. Over time, consistent use of an improvement rate strategy can eventually pull student performance into acceptable ranges. This strategy is preferable for programs that originally produce lower quality results.

CAN STUDENT FEEDBACK BE USED TO IMPROVE OVERALL ASSESSMENT PRACTICE?

To improve the learning gains from any assessment or learning activity, debriefing will be helpful to both student and teacher. Students can provide feedback either before or after testing that can influence changes that might need to be made. Requiring that students discuss the impact of the testing will improve their retention related to the assessment.

Another strategy that particularly helps involves student feedback. At the conclusion of a course, ask students which activities and tests should be kept, which should be revised, and which should be replaced for improvement in helping students meet their goals. Ask each student to pick one in each category, compile the results, and make recommended refinements.

5. USE ASSESSMENT DATA EFFECTIVELY AND ETHICALLY

The final step in the assessment process involves taking action based on the results, particularly in relation to making a judgment about whether both the teaching and learning that transpired are sufficient. In this section, we explore what it means to "close the loop" on assessment. On a practical level, this decision means that the results reflect one of the following conditions:

- The results are satisfying; the students have learned what was expected and, conversely, you have taught them well.
- Their results are disappointing; the flaw may lie with incompetent responses by the students, ineffective teaching, and/or problematic assessment design.

Regardless of the level of achievement, most assessment strategies provide feedback about possibilities for improvement. Next we consider each of these scenarios with regard to the use of assessment results.

WHAT ARE MY ETHICAL OBLIGATIONS RELATED TO ASSESSMENT OF STUDENTS?

Ethical behavior is involved in preparing and executing the assessment as well as in processing the results.

Before
Be sure to design the assessment with the idea of making the assessment a level playing field. Students should be assessed on content and skills directly related to the course. Wherever possible, you should strive to rule out inflating performance due to privilege or other facts not specifically related to the course (i.e., make sure test items don't rely on cultural references unfamiliar to some of your students).

During
Assessment conditions should foster optimal performance. Provide clear instructions and a quiet environment. Be on hand to answer questions for clarification. Encourage students to do their best work.

After
Score performances quickly to maximize learning from the experience. Construct respectful and encouraging feedback referencing student successes and what they can do to improve their performance. Student performance is a private matter. Exercise caution regarding how you describe assessment results in public. Individual identity should remain anonymous.

HOW DO I EXPLOIT SUCCESSFUL RESULTS?

Positive assessment results have enormous value. Good assessment results serve as evidence of high-quality teaching and learning and can be used to advocate for new resources in the program (e.g., field trip funding). Strong evidence of learning can also assist you in making the case that you are doing your job well.

When your results are successful, you can declare a new "maintenance" goal and continue the assessment with future students. Your objective is ensuring that the solid achievement remains sturdy over time. Especially when an assessment strategy measures some central feature of the curriculum, it is easy to justify repeating the same goals in future terms. However, a different institutional assessment strategy is to explore a specific assessment question (e.g., "How well do my students write?"). When assessment provides an answer to that question, the program may be expected to move on to a new assessment question.

WHAT IF MY RESULTS ARE DISAPPOINTING?

You have to decide where the problem lies. Disappointing results are not necessarily "bad" results, since the evidence can point to real problems that need to be addressed. An analysis of performance weaknesses can dictate what might need to be changed.

The learning factor
An assessment that reveals disappointing performance may accurately capture the gaps in student knowledge and skill. Instructors need to determine whether the problem rests with inadequate exposure, insufficient practice, fatigue, distraction, unmotivated performance, or any number of student variables that could contribute to lackluster performance.

The teaching factor
Poor results can indicate that the problem lies with the teacher. Perhaps the teaching was rushed or pitched at too high a level for the students to grasp. For example, suboptimal performances in writing might inspire the instructor to think of other strategies that could build that skill to produce stronger performances the next time the assessment transpires.

The assessment design factor
Poor performance can also be an indicator that the assessment is not well designed or may not be a good match for the intended goals. Assessment is an iterative process; it is normal for procedures to undergo continuous refinement to improve confidence in its use. The "low tech" and "high tech" suggestions mentioned earlier are helpful tools teachers can use to diagnose possible problems and revise assessments.

HOW SHOULD I REPORT MY RESULTS?

Most education programs have developed vehicles for reporting assessment results. Typically, a school specifies a strategy for collecting the conclusions about assessment results that facilitate comparisons of success across units within the organization. Follow the reporting requirements, but take advantage of any opportunities to include easy-to-understand attachments or graphs that readily communicate your success. It is also helpful to create a one-page executive summary to capture high-quality achievement or articulate where your assessment planning will take you. For example, teachers could use this kind of executive summary to share overall test results with students, modeling effective data presentation and providing potentially useful overall suggestions about common misconceptions. Websites can also showcase assessment success.

HOW CAN I PROTECT MYSELF IF THE RESULTS ARE DISAPPOINTING?

If teachers are to take assessment seriously, assessment should not be threatening. The results of assessment should not be grounds for punitive action taken by administration. Such conditions are likely to give rise to problematic execution and even misrepresentation of results. However, if administrations provide evidence that assessment is underway in the spirit of continuous improvement, teachers should have greater motivation to use assessment as the powerful tool that it is to promote student learning and improve teaching. Technically, administrations should support teacher assessment activities as long as teachers are making legitimate efforts to gather relevant data and work with those data to help them determine teaching strategy.

If results are problematic, then as a matter of ethics they should not be ignored, downplayed, or distorted. However, the following steps can mitigate negative consequences from disappointing student outcome performance.

- Lead with the strengths observed in the assessment.
- Fine-tune goals based on what the data reveal.
- Be clear that actions will be taken to address the deficits. Promise better outcomes the following year and then deliver.
- Work with a mentor whose outcome performance is satisfying to help you craft an improvement strategy.

REFERENCES

American Psychological Association. (2011). *National standards for high school psychology curricula.* Retrieved from http://www.apa.org/education/k12/national-standards.aspx

American Psychological Association, Working Group on Introductory Psychology Assessment. (2017). *Assessment of outcomes of the introductory course in psychology.* Retrieved from https://www.apa.org/ed/precollege/assessment-outcomes.pdf

Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing: *A revision of Bloom's taxonomy of educational objectives.* New York, NY: Longman.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: Vol. 1: Cognitive domain.* New York, NY: McKay.

Brookhart, S. M. (2017, May). *Assessing what students know and can do.* Paper presented at ASCD Empower17: The Conference for Every Educator, Anaheim, CA.

Gonzalez, J. (2014, Oct. 9). *Your rubric is a hot mess: Here's how to fix it* [Blog post]. Retrieved from http://www.brilliant-insane.com/2014/10/single-point-rubric.html

Gramoll, K. (n.d.) *eCourses: Grading.* Retrieved from https://www.ecourses.ou.edu/cgi bin/info.cgi?topic=grading

Karbach, M. (2014, Feb. 5). *A visual chart on summative vs. formative assessment* [Blog post]. Retrieved from http://www.educatorstechnology.com/2014/02/a-visual-chart-on-summative-vs.html

Reiner C. M. (2002). *Preparing effective essay questions: A self-directed workbook for educators.* Retrieved from https://testing.byu.edu/handbooks/WritingEffectiveEssayQuestions.pdf

Webb, N. (1997). *Research Monograph Number 6: Criteria for alignment of expectations and assessments on mathematics and science education.* Washington, DC:Council of Chief State School Officers. Retrieved from https://eric.ed.gov/?id=ED414305

Webb leveling: *Expectations for student performance.* Retrieved from http://doe.sd.gov/octe documents/WebbLevel.pdf

White, J. (2018, April 2). T*echnology Tips & Tricks: Assessment and Engagement* [Blog post]. Retrieved from http://psychlearningcurve.org/technology-tips-tricks-assessment-engagement/

Wilson, L. O. (2016). A*nderson and Krathwohl—Bloom's taxonomy revised.* Retrieved from http://thesecondprinciple.com/teaching-essentials/beyond-bloom-cognitive-taxonomy-revised/

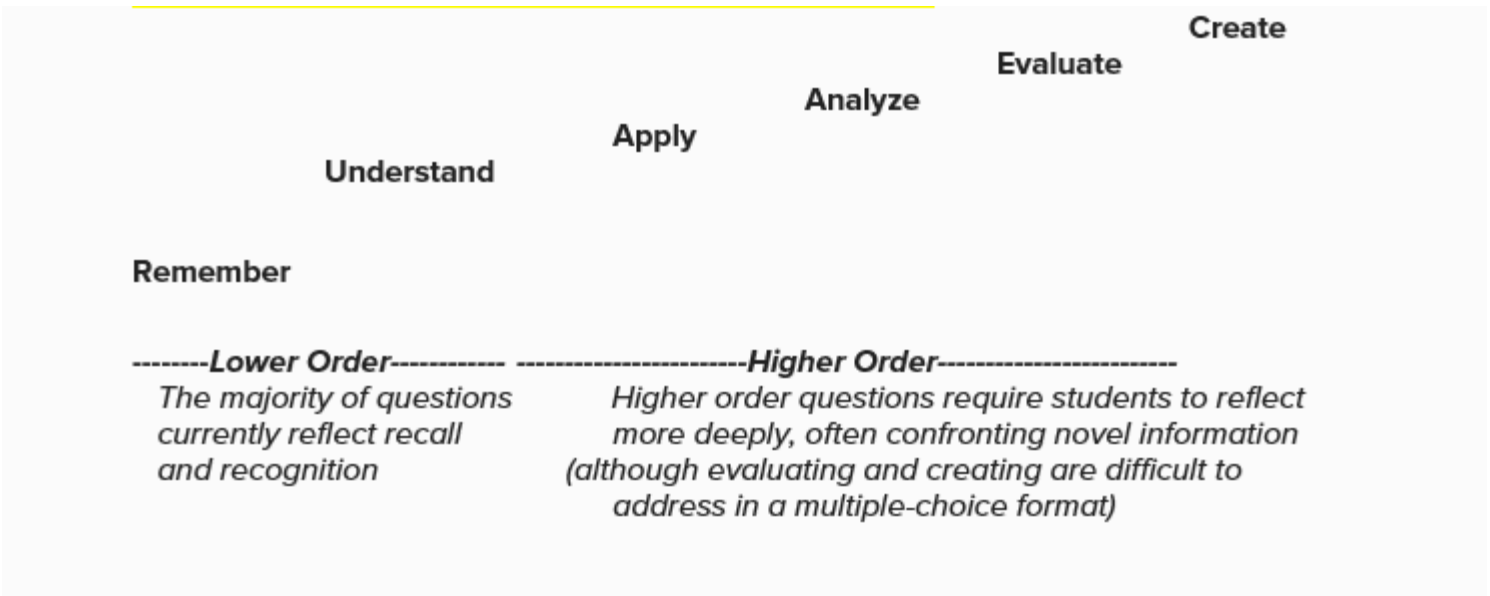
APPENDIX A

A Guide to Writing Higher Order Multiple-Choice Questions for the AP Psychology Test

The purpose of this guide is to assist item writers in developing more challenging test questions. In the tradition of Bloom's Taxonomy, we refer to those questions as "higher order." We offer some pointers and examples to facilitate the development of such questions.

PROMOTING COGNITIVE COMPLEXITY IN APPLYING BLOOM'S TAXONOMY

Bloom's Taxonomy (1956) presents a way to think about increasing levels of cognitive complexity. Krathwohl and colleagues (2001) modified Bloom's original taxonomy to correct for some operational difficulties of the original taxonomy. The modified taxonomy is as follows.



CREATING LOWER ORDER QUESTIONS

Lower order questions tend to reflect two kinds of structures. The concept or principle is in the stem of the question that has a number of potential answers (the examples we offer have five alternatives). The correct answer (denoted with an asterisk in the examples) represents a clear attribute of the concept or principle.

Obsessive-compulsive disorder is characterized by which primary symptom?

1. Hallucination
2. Memory loss
3. Intense specific fear
4. Delusion
5. Unwanted repetitive thoughts*

The second format places the attribute in the stem of the question; the correct alternative is the target concept or principle.

Which disorder is characterized by unwanted, intrusive thoughts and repetitive behavior?

1. Phobia
2. Obsessive-compulsive disorder*
3. Dissociative identity disorder
4. Major depressive disorder
5. Schizophrenia

A third example offers a slightly more challenging lower level question:

Jane checks 10 times to see whether she has set her alarm clock before going to sleep. This behavior may be symptomatic of which of the following disorders?

1. Phobia
2. Obsessive-compulsive disorder*
3. Dissociative identity disorder
4. Major depressive disorder
5. Schizophrenia

When creating lower order questions, question writers should

- identify a concept that is meaningful and discussed by the majority of textbooks;
- create real and plausible alternatives that reflect the same category of responses;
- clearly distinguish the correct alternative; and
- use key verbs such as recognize, recall, interpret, classify, summarize, infer, compare, explain.

CREATING HIGHER ORDER QUESTIONS

The key to creating higher order questions is requiring students to mediate their answers by coming up with an extra step that they had not previously learned in their studies. It is not always easy to distinguish application questions from analysis questions. In general, application and analysis questions ask students to transfer their recalled knowledge to a new situation, break apart and reassemble concepts in new ways, or combine the content of two areas in a novel way to answer a question.

Application questions

A student who misses deadlines in school while striving for perfection may be exhibiting the symptoms of which of the following disorders?

1. Phobia
2. Obsessive-compulsive disorder*
3. Dissociative identity disorder
4. Major depressive disorder
5. Schizophrenia

Which of the following could be expected to be a problem for a student with obsessive-compulsive disorder?

1. Cheating without remorse
2. Oversleeping and missing class
3. Missing deadlines while striving for perfection*
4. Alienating classmates due to inflated self-importance
5. Being distracted by fear of persecution by classmates

Analysis questions

The following examples involve combining content from two or more areas of psychology, which forces the student to consider concepts in a novel manner:

Gene is always late for school because he spends an hour organizing his closet each morning. Which of the following treatments would be most effective for Gene's problem?

1. In-depth interpretation of dreams
2. Electroconvulsive therapy
3. Medication affecting serotonin levels*
4. Systematic desensitization
5. Regular exposure to bright lights

Amanda washes her hands until they bleed. This causes her distress, but she still cannot stop washing her hands. A behaviorist would explain the persistence of Amanda's behavior as a result of

1. positive reinforcement due to increased anxiety*
2. the extinction of her interest in other activities
3. generalization from taking long showers
4. negative reinforcement due to decreased anxiety
5. time out from other meaningful activities

When creating higher order questions, question writers should

- pose a problem that needs to be solved;
- consider eliciting interpretation of a chart, graph, or picture;
- ensure that situations described in the question are novel; and
- use keywords such as execute, organize, apply.

REFERENCES

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* New York, NY: Longman.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: Vol. 1: Cognitive domain.* New York, NY: McKay.

APPENDIX B

Evaluating Assessment Strategies

The group of educators who met at the APA Summit on National Assessment of Psychology (2016) developed a rubric to help instructors make good decisions about whether an instrument will meet their needs. Based on the works of Fulks (2004), Rhode Island Department of Education (2012), and CAST (n.d.), this is a tool that may be used to evaluate assessments. The tool includes an assessment profile (e.g., cost of instrument, time for implementation, cognitive taxonomy), supporting information (e.g., teacher directions, student directions), delivery method (e.g., whole group, individual, electronic), how it maps onto the APA Guidelines for the Undergraduate Psychology Major and Universal Design for Learning Guidelines (see CAST, n.d.), a scoring guide for rubrics, and evaluation of strengths and weaknesses for assessment.

Assessment Profile	Response
Method (I = Indirect and D = Direct)	
Level of Assessment (SaS = Students as Scholars, SaSC = Students as Savvy Consumers, and SaP = Students as Producers)	
Cognitive Taxonomy (R = Remembering, U = Understanding, A = Applying, AN = Analyzing, E = Evaluating, C = Creating	
Usage (F = Formative and S = Summative)	
Constructed Response (essay, multistep response with explanation and/or rationale required for tasks, etc.)	
Product (research paper, editorial, log, journal, play, poem, model, multimedia, art products, script, musical score, portfolio pieces, etc.)	
Performance (demonstration, presentation, science lab, dance or music performance, athletic performance, debate, etc.)	
Short Answer (short constructed response, fill in a graphic organizer or diagram, explain your thinking or solution, make and complete a table, etc.)	
Selected Response (multiple choice, multiple select, evidence-based selected response (EBSR), true–false, matching, etc.)	
Cost	
Time (in minutes)	

Supporting Information	Yes/No
Teacher Directions (may include prerequisites/description of instruction before giving the assessment; e.g., This assessment should be given after students have learned...)	
Scoring Guide/Rubric	
Sample Evidence for Student Performance	
Student Materials	
Estimated Time for Administration	
Student Directions	
Assessment Task/Prompt	

Delivery Method	Yes/No
Whole Group	
Small Group	
Individual	
Paper and Pencil	
Electronic	

APA Guidelines for the Undergraduate Psychology Major 2.0 Evaluation	Fully	Partially	N/A
1. Knowledge Base in Psychology			
2. Scientific Inquiry and Critical Thinking			
3. Ethical and Social Responsibility in a Diverse World			
4. Communication			
5. Professional Development			
Describe the content knowledge/concepts assessed:			
Describe the skills/performance assessed:			

Scoring Guide to Be Used With the Assessment:	Yes/No
Generalized Rubric (e.g., for writing an argument, for all science labs)	
Task-Specific Rubric (only used for the particular task)	
Scoring Guidelines (e.g., checklist with score points for each part)	
Answer Key, Scoring Template, Computerized or Machine Scored	
Anchor Papers (student samples at each score point)	
Are the score categories clearly defined and coherent across performance levels? If no, please explain.	
Does the rubric/scoring criteria address all of the demands within the task or item? If no, please explain.	
Are directions for the items or tasks presented in as straightforward a manner as possible for a range of learners? If no, identify problematic items/tasks and provide suggestions for improvement.	
Is the vocabulary and context(s) presented free from cultural or other unintended bias? If no, identify problematic items/tasks and provide suggestions for improvement.	

Universal Design for Learning	Yes/No
Provide multiple means of accessing the assessment (e.g., allow students to access information in ways that do not require them to visually read standard print).	
Provide multiple means of responding to the assessment (e.g., allow students to complete activities, assignments, and assessments in different ways or to solve or organize problems using some type of assistive device or organizer).	

Recommendations for This Assessment:
What are the strengths of this assessment?
What are the weaknesses of this assessment?
☐ This assessment can be used without revisions.
☐ This assessment can be used with minor revisions (explain below)
☐ This assessment can be used with significant revisions (explain below)
☐ This assessment should not be used (explain below)

REFERENCES

CAST. (n.d.). *About universal design for learning*. Retrieved from http://www.cast.org ourwork/about-udl.html#.WobYieyWy5t

Fulks, J. (2004). *Assessing student learning in community colleges*. Retrieved from http://www2.bakersfieldcollege.edu/courseassessment/

Rhode Island Department of Education and National Center for the Improvement of Educational Assessment. (2012). *Assessment review tool: A part of the assessment toolkit* (Revised 2012). Retrieved from http://www.ride.ri.gov/Portals/0/Uploads/Documents/Teachers-and-Administrators-Excellent-Educators/Educator-Evaluation/Online-Modules/Assessment-Review-Tool.pdf