

APPENDIX 3B
EVALUATION OF STUDENT ACHIEVEMENT
GUIDELINES FOR EVALUATION INSTRUMENTS AND ITEMS

CONTENTS:

WRITTEN TEST ITEMS

Selected-Response Items

 Matching Test Items

 Multiple-Choice Test Items

 True-False Test Items

Constructed-Response Items

 Listing Test Items

 Completion Test Items

 Short-Answer Test Items

 Essay Test Items

PROCESS AND PRODUCT EVALUATIONS

Performance Checklists

Product Evaluation Forms

Rating Scales

Product Evaluation Guidelines

Interrater Reliability

WRITTEN TEST ITEMS: SELECTED RESPONSE ITEMS

Matching Test Items

The standard matching format consists of two lists containing related words, phrases, or symbols. The student is required to match elements on one list with associated elements on the other list according to specific instructions. The student pairs the elements in each list through logical association and records the answer. Generally, a matching test item is used when it is difficult to develop plausible distracters for individual items (e.g., terms and definitions) and the items are closely related.

In the guidelines below, "stimulus items/entries" are the items to be matched, normally included in the first column. Each stimulus item is usually preceded by a blank where students record their selection from the response column or columns.

Guidelines

1. The directions should include:
 - a. An explanation of what is in each column.
 - b. A statement of how entries in the columns should be matched.
 - c. How often choices in the response column(s) may be used.
 - d. How many responses are possible for each stimulus item.
2. Items in each list should be homogeneous. Dissimilar items are too easily matched; closely related items provide a better discrimination of knowledge.
3. The matching exercise should include at least four stimulus items to decrease the effectiveness of guessing.
4. No more than two response columns should be used. A single response column is preferable.
5. Plausible distracters (response items that will not be used) should be included in the response column(s) to decrease the effectiveness of guessing.

CDG (Tech) - App 3B

6. Response items should be arranged in some logical order (e.g., ascending or descending numeric order or alphabetically) whenever possible.
7. Both stimulus and response items should be as short as possible.

Multiple-choice Test Items

Multiple-choice test items consist of two parts. The problem statement is presented in the stem, followed by a set of alternatives. The correct response is called the key. Incorrect alternatives are called distracters. Typically either four or five possible responses are included in the set of alternatives. With four alternatives, students have a better chance of guessing the correct response (1 out of 4 vice 1 out of 5), but it may be difficult to come up with four plausible distracters.

Guidelines

1. Wording in the stem must be clear and unambiguous, so that only one answer is correct.
2. Negatives generally should not be used. Negative wording makes items harder to read. Students may miss the item because they misread it rather than because they did not know the content. If a negative must be used in the stem or alternatives, it must be highlighted (e.g., bold face or italic script) for emphasis.
3. Distracters must be plausible. "Throw away" alternatives should not be included.
4. Distracters should be based on common misconceptions or errors whenever possible.
5. Each distracter must be demonstrably **incorrect**. Including distracters that are "less correct" than the key **is not acceptable**. The stem may require students to select the "most likely" or "best" response in a set of alternatives when answers **not included in the set of alternatives** would be correct. For

example, students may be asked to select the most likely interpretation of a set of data from the set of alternatives. The key must be a likely interpretation of the data. Each distracter must be an incorrect interpretation of the data. "Most likely interpretation" or "best interpretation" could be used in the stem because possibilities exist outside of the given set of alternatives.

6. Distracters such as "all of the above," "none of the above," and "a and b above" should be avoided. Such alternatives increase the effectiveness of guessing. Similarly, avoid "multiple multiple-choice" questions that require students to select combinations of alternatives (e.g., "K questions").

7. All repetitive phrases must be included in the stem.

8. All alternatives must be grammatically consistent with the stem. Alternatives that are not grammatically consistent with the stem are usually incorrect and can be easily eliminated by students.

9. Alternatives should be arranged in some regular manner. Numeric alternatives may be arranged in ascending or descending order. Verbal alternatives can be arranged from shortest to longest if they vary much in length or alphabetically if they are close to the same length. This tends to randomize the placement of the key. NOTE: If verbal alternatives are arranged by length, avoid having the correct answer always the longest or shortest alternative.

True-false Test Items

True-false test items present a statement to the student who must then determine whether or not the statement is correct. It essentially represents a two-response multiple-choice item. Students therefore have a 50-50 chance of guessing the correct answers.

Guidelines

1. The statement must be concise and clear. The proposition which is to be judged as true or false must be evident.

CDG (Tech) - App 3B

2. The statement must be clearly true or false; avoid shades of meaning.
3. Only a single idea or piece of information should be tested in one item.

If combinations are absolutely necessary, students must be given clear directions on how to classify partially true or partially false statements (i.e., "A statement is false if any part of the statement is false.").

4. A false statement must be consistent with a typical misconception.

5. The factor that makes a false statement false must be significant. For example, a true-false test item in the form "According to the Curriculum Development Guide," the term "soft skills" is defined as "_____", would be true as long as the definition provided is a correct paraphrase of the definition in the reference. Students should not be expected to distinguish verbatim from paraphrased information.

6. Qualifiers such as always, never, none, all, may, or sometimes should be avoided. Items that include words such as "never" or "always" are usually false because there are exceptions to most rules. Items that include words such as "sometimes" or "may" are usually true.

WRITTEN TEST ITEMS: CONSTRUCTED-RESPONSE ITEMS

Listing Test Items

In listing test items, students are asked to generate a list of some kind, such as a list of the steps in a procedure.

Guidelines

1. Listing test items are normally phrased as directives (e.g., "List the steps in the procedure to ...").
2. Directions for each item should include:
 - a. The type of information that should be included in the list (e.g., "List the periodic maintenance that must be completed on the XYZ 2000. Daily, weekly, monthly, quarterly, and yearly requirements must be included.").

b. Any elaboration required (e.g., "List and describe the quarterly maintenance procedures that must be completed on the XYZ 2000 to maintain the warranty.").

c. The type of sequence required, if any (e.g., "List the steps required to change the dindlehopper on the XYZ 2000. Steps must be listed in the order they would be performed.").

d. Whether or not spelling and/or exact wording is required.

3. The scoring key for each item should include:

a. Any alternatives or synonyms that may be included.

b. How the total number of points will be distributed if partial credit is allowed. This includes considerations such as points lost for incorrect spelling as well as partial credit for incomplete answers.

c. How sequence will be graded if a sequence is specified in the directions (NOTE: sequence and completeness should be scored separately).

Completion Test Items

In completion test items, students are presented with a statement including a blank that must be filled in. Completion test items are also called "fill-in-the-blank" test items because of this structure.

Guidelines

1. Items should include only one blank. More than one blank often results in an item with little meaning. For example, in the item "The symbol for _____ is _____," any pair containing a symbol and what it represents is correct.

2. The information to be filled in must be significant. Only things that are important for the student to remember are left blank.

3. Items should be phrased so that only one word or phrase will complete the sentence correctly. For example, "Columbus discovered America in

CDG (Tech) - App 3B

_____, " could be correctly completed with "the Santa Maria." If the item was intended to elicit the year of discovery, it should say so. A better wording would be "Columbus discovered America in the year ____."

4. Grammatical cues to the correct answer must be avoided. For example, in the item "A completion item should have at most _____ blank," the only reasonable answer is "one" because the word "blank" is singular. Using "blank(s)" instead of "blank" would eliminate this problem. Similarly, "a(n)" should be used instead of "a."

5. The blank should be placed near the end of the item.

6. Directions should indicate when correct spelling or exact wording is required.

7. All acceptable synonyms or alternative correct answers must be specified in the scoring key.

Short-answer Test Items

In short-answer test items, the student is expected to construct a short response to a question. The response is usually no more than one or two sentences and may be a single word or phrase.

Guidelines

1. Short-answer items may be phrased either as questions (i.e., "What is the definition of...") or as directives (i.e., "State the definition of...").

2. The question/directive should be clear and complete. It should tell the student clearly what information should be included in the answer.

3. The required answers should be short; no longer than one or two sentences.

4. The directions should include how the items will be scored without giving cues to the correct answer. They should also indicate when correct spelling or exact wording is required.

5. The scoring key should include all acceptable synonyms, alternatives, or forms of the correct answer. If partial credit can be awarded, the number of points allowed for each factor in the expected response should be specified.

Essay Test Items

Essay test items, or essay questions, require the student to develop and organize a response to a relatively broad question. Responses to essay items are normally at least a paragraph in length.

Guidelines

1. Essay items are normally phrased as directives (i.e., "Explain the rationale for each of the principles of asepsis established by the AAORN.").
2. The question/directive should be clear and complete. It should tell the student clearly what information should be included in the answer.
4. The directions should include how the items will be scored without giving cues to the correct answer.
5. The scoring key should include all of the factors or information students should include in their response. The scoring key should also indicate how credit will be awarded for the item. If presentation of the response will be graded as well as content, two scoring keys are needed; one for accuracy and completeness (i.e., the factors that need to be included) and one for other considerations (e.g., grammar, spelling, logical development of an argument, neatness).

PROCESS AND PRODUCT EVALUATIONS

Performance Checklists

Performance checklists (PCLs) are assessment tools used to evaluate satisfactory performance and reduce evaluation subjectivity for learning

CDG (Tech) - App 3B

objectives that require the student to demonstrate specified skills, tasks, or procedures.

Each PCL contains all the steps that must be satisfactorily performed for the student to be successful. Safety precautions and use of instruments, supplies, tools, equipment, or facilities can be included in PCLs as well as procedural steps to be evaluated. Steps may also be included that measure students' understanding of the skill, task, or procedure to be demonstrated. Such steps could require students to explain the purpose of the skill, task, or procedure; what they will do next; or why a step is performed in a certain way.

Most PCLs contain only two response options. Instructor/proctor observation is required and each step of the PCL is checked satisfactory or unsatisfactory, pass or fail.

Guidelines

Each PCL should include the following sections:

1. PCL name and number: Each PCL will bear the number of the terminal or enabling objective it evaluates.
2. Date the PCL was developed or revised, placed in the upper right hand corner of the first page of the PCL.
3. Block or line for student's name and the date of the evaluation.
4. Block or line for performance score (e.g., pass or fail; satisfactory or unsatisfactory; numerical score).
5. Objective and objective number: The objective and its number will be stated as in the curriculum outline.
6. Reference(s) for development: The source or sources used in developing the PCL (i.e., the document(s) that describe "correct performance") should be listed.
7. Requirements for successful completion: Specify exactly what the student must do in order to successfully complete the PCL. For example, "all

steps must be performed satisfactorily;" "all critical steps and at least 3 of the 5 non-critical steps must be performed satisfactorily." Specify any series of steps that must be performed in sequence.

8. Directions for scoring: Directions should specify exactly how the checklist is to be used and scored to ensure reliability of results if there is more than one evaluator. If a numerical grade is to be assigned, the directions must specify how the grade is computed.

9. Elements (or steps) to be evaluated: The skill, task, or procedure to be demonstrated must be broken down into its important elements or steps. The breakdown should be based on the steps an expert would perform and where, when, and how he/she would perform those steps. For example, if a particular step of a task is to "place a chart in front of patient's face" (for testing visual acuity) and it is important that the chart be placed 14" to 16" from the patient's face, the step should be phrased "place chart 14" to 16" from patient's face." Each step should be carefully worded to avoid disagreement about the meaning of the statement or about the judgement that the step was or was not correctly performed.

10. Designation of critical elements: A step is considered a critical element if the health or safety of the patient, the student performing the step, or other staff members would be endangered if the step is not performed correctly; or if incorrect performance would damage expensive equipment. Critical elements should be marked with an asterisk.

11. Remarks section: If the student's performance on any step is marked "unsatisfactory" the instructor must include feedback with an explanation of the unsatisfactory mark to the student in the remarks section. Feedback may also include suggestions for improving skills, references to read, and the date of the next evaluation if the student must be retested.

CDG (Tech) - App 3B

12. Block for student's signature and date: After the evaluation has been completed and reviewed with the student, the student should sign and date the PCL to acknowledge that the instructor has reviewed the results with him/her.

13. Block for instructor's signature: The instructor signs the PCL upon completion of the evaluation and review with the student.

Product Evaluation Forms

Product evaluation forms (PEFs) are assessment tools to evaluate products produced by students when the products from all students will be evaluated on the basis of standard physical characteristics. For example, if a learning objective requires students to fabricate a custom impression tray, the important characteristics that describe what the finished product should or should not look like would be included on the PEF.

Guidelines

Each PEF should include the following sections:

1. PEF name and number: Each PEF will bear the number of the terminal or enabling objective it evaluates.
2. Date the PEF was developed or revised, placed in the upper right hand corner of the first page of the PEF.
3. Block or line for student's name and the date of the evaluation.
4. Block or line for performance score (e.g., pass or fail; satisfactory or unsatisfactory; numerical score).
5. Objective and objective number: The objective and its number will be stated as in the curriculum outline.
6. Reference(s) for development: The source or sources used in developing the PEF (i.e., the document(s) that describe the characteristics that must or must not be present in an acceptable product) should be listed.

7. Requirements for successful completion: Specify exactly what the student must do in order to successfully complete the PEF. For example, "all steps must be performed satisfactorily;" "all critical steps and at least 3 of the 5 non-critical steps must be performed satisfactorily." Specify any series of steps that must be performed in sequence.

8. Directions for scoring: Directions should specify exactly how the form is to be used and scored to ensure reliability of results if there is more than one evaluator. If a numerical grade is to be assigned, the directions must specify how the grade is computed.

9. Elements (or characteristics) to be evaluated: Collect samples of the product and sort them into good and bad categories. Go through all the good samples and note the characteristics they have in common that make them good. Go through all the bad samples and note characteristics that make them bad. Then refine the good and bad characteristics in consultation with job experts. The result is a list of characteristics of the product. State the characteristics positively (i.e., as characteristics that should be present).

10. Designation of critical elements: Any element that would make the product irretrievably unacceptable (i.e., the product cannot be used even with modifications) may be considered critical. Any elements considered critical should be marked with an asterisk.

11. Remarks section: If any characteristic is marked "unsatisfactory," the instructor must include feedback with an explanation of the unsatisfactory mark to the student in the remarks section. Feedback may also include suggestions for improving the product, references to read, and the date of the next evaluation if the student must be retested.

12. Block for student's signature and date: After the evaluation has been completed and reviewed with the student, the student should sign and date the PEF to acknowledge that the instructor has reviewed the results with him/her.

CDG (Tech) - App 3B

13. Block for instructor's signature: The instructor signs the PEF upon completion of the evaluation and review with the student.

Rating Scales

Rating scales are adjuncts to PCLs or PEFs that allow evaluation of levels of performance for a process, product, or both. Rating scales may be incorporated on a PCL or PEF if a product or step in a procedure may vary in qualities that can be objectively defined or described at different levels. A rating scale has diagnostic value in that the evaluator and student can tell exactly where improvement may be made.

Guidelines

1. Only elements that have distinct levels of satisfactory or unsatisfactory characteristics or behaviors may be evaluated using rating scales. Therefore, a rating scale may be incorporated into a PCL/PEF for some steps/characteristics and not others.

2. For each element (step or characteristic) to be rated, state the quality associated with each successive level.

3. Each level should be discrete and readily differentiated from all other levels. Normally, this will require unique descriptions for each level of each element to be rated.

4. Each rating scale should evaluate one factor. For example, if speed and accuracy are both ratable factors in performing a particular step in a process, separate rating scales should be developed for the speed factor and for the accuracy factor.

5. Global designators (e.g., outstanding, excellent, marginal) applied across elements are not acceptable.

6. Unacceptable levels of performance must be clearly identified.

7. Rating scales should be reviewed with subject matter experts to verify accuracy (i.e., that elements are correctly defined) and clarity (i.e., that levels are adequately defined).

8. When rating scales are incorporated in PCLs or PEFs, the requirements for successful completion and directions for scoring should include reference to the rating scales.

Product Evaluation Guidelines

Product evaluation guidelines are used when a learning objective requires students to produce a product, but the evaluation of the product does not depend on standard physical characteristics or the products developed by students will not be uniform (e.g., student projects where individuals may select different types of projects). Product evaluation guidelines are normally used for written products.

Product evaluation guidelines serve two purposes. First, they minimize instructor subjectivity in grading the products. Secondly, they inform students of the parts and characteristics that are required in the product and the relative importance of each part and/or characteristic.

Guidelines

1. The minimum criteria for each part or section of the product must be described. For example, if students are required to produce case worksheets (as in the Surgical Technologist training program), the minimum criteria for each section of the case worksheet must be described.

2. The factors that will be evaluated in the product as a whole and in each part/section must be stated. For example, if errors in grammar or spelling will impact on the grade for the product, that must be specified.

3. Scoring directions must clearly establish the weighting assigned to each factor and section/part.

CDG (Tech) - App 3B

4. Products may be graded on a pass/fail basis or a numeric or letter grade may be assigned.

Interrater Reliability

Establishing interrater reliability is an important aspect in determining the validity of evaluation tools such as performance checklists and product evaluation forms. Performance checklists and product evaluation forms are used to minimize subjectivity in evaluating student demonstrations or products. If the evaluation tool used is reliable and instructors have been trained in its use, variability among raters will be minimal.

The key to establishing interrater reliability is to compare ratings assigned to a single performance or product by a number of raters using the same evaluation tool. Three to five raters or evaluators should participate in this endeavor. If possible, multiple performances or products should also be used to compare the ratings assigned by individual raters to different performances or products. This will provide a better sample of ratings.

It is recommended that the performance(s) to be evaluated be videotaped. If multiple performances will be used, they should vary from unacceptable to highly proficient in terms of student competency. This is particularly true if rating scales are incorporated in the performance checklist. Performances by different students are preferable, if that can be arranged; or instructors could be videotaped demonstrating the performance including predetermined errors frequently made by students.

For each performance, raters complete the checklist as if grading students in the normal training setting. The results from all raters across performances are then compared. If multiple performances are evaluated, the results across performances by each rater are also compared. A chart similar to the one below can be used to compare ratings across raters. This chart reflects

the rating assigned to five items using a 1 to 5 rating scale. Four raters evaluated four performances.

SAMPLE RATING SCALE DATA
4 Performances, 4 Raters, 5 Steps Rated

	Performance 1				Performance 2				Performance 3				Performance 4			
	Raters				Raters				Raters				Raters			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
S1	5	5	5	5	3	3	3	3	4	4	4	4	3	3	3	3
S2	5	4	4	5	4	4	4	4	3	4	4	4	1	2	2	2
S3	5	4	5	4	4	3	3	4	3	3	3	3	3	2	2	2
S4	3	5	2	4	3	1	4	2	2	4	3	2	1	2	4	3
S5	4	4	4	3	3	2	3	2	4	3	3	3	3	2	2	2

R1=Rater 1, R2=Rater 2, R3=Rater 3, R4=Rater 4
 S1=Step 1, S2=Step 2, S3=Step 3, S4=Step 4, S5=Step5

The scores given to each performance by the different raters can be compared by looking across each row. In this example, there is perfect agreement among raters for step 1. For steps 2, 3, and 5, there is some disagreement and for step 4, there is considerable disagreement. For all but step 4, the interrater reliability is probably acceptable if all of the ratings fall within the criteria for passing (disagreement between raters on acceptable and unacceptable performance is covered on the next page). The majority of the decisions are the same, and the range is slight (one level up or down) where they do differ. In general, if there is no majority agreement or if raters differ by two or more points on the scale, the step should be reviewed.

This would be the case with step 4. The levels for the rating scales for step 4 may need to be clarified and the instructions to scorers should be checked to make sure they are not misleading some of the raters. It is best to

CDG (Tech) - App 3B

do this with the raters, because they can explain what they intended. If the performances were videotaped, it may be easier for raters to recall and explain why they assigned the particular rating to a step when they review the videotape.

The agreement on decisions of pass or fail needs to be tighter (the consequences of disagreeing on a pass/fail decision are greater than on an acceptable/better decision). If a rating of "1" in the example above reflects unsatisfactory performance, and "2" and above reflect satisfactory performance, the differences in rating for step 2 in the fourth performance would need to be investigated. With a pass/fail grading system, any disagreement between raters should be investigated.