Excerpts from:

Rodriguez, M.C. (2016). Selected-response item development. In S. Lane, M. Raymond, & T.M. Haladyna (Eds.), *Handbook of test development* (2nd ed., 259-273). New York, NY: Routledge.

Choosing the SR Item Format

For most testing purposes, the test developer must choose between SR and constructed-response (CR) formats. The advantages and disadvantages of SR items, particularly MC items, have been reviewed in comparison to CR items (Rodriguez, 2002, 2003)1. Many of the SR formats available to the test developer are described below. Similarly, there are many forms of CR items available, most commonly including short-answer or extended response items, or other formats where test takers must generate a response (grid-in items, graphical manipulation) rather than select one from a set of options. Among the advantages, SR items support:

- Direct measurement of many skills, including abilities to discriminate, to understand concepts and principles, make judgments about next steps, draw inferences, reason about arguments, complete statements, interpret data, apply information;
- Administration efficiency and objectivity of scoring;
- Response efficiency – not requiring students to write;
- Potential for diagnostic information from distractor analysis;
- Broad sampling of content domain.

In comparison, CR items are more appropriate when the target of measurement requires a written response, when novel solutions are desired, complex process information is needed through synthesizing, organizing, and sequencing information, or explanations are required. This includes a wide range of performance assessments beyond the focus of this chapter, including scenarios where subject matter experts create tasks that mimic actual procedures in the field (target domain).

The disadvantages of SR formats are just as onerous:

- Indirect assessment of some skills, such as ability to recall or explain concepts, provide or express ideas, organize or construct something;
- The fixed options limit the expression of ideas or novel solutions;
- Knowledge may appear to be artificially constructed – absent real-world contexts;
- Reading skills may interfere in the assessment of knowledge and skills in other domains
- May be susceptible to guessing.

Again, in comparison, CR items also have limitations. Scoring can be a challenge when there are multiple logical or acceptable responses and when the responses involve novel solution strategies – resulting in lower score reliability because of scoring subjectivity. Short-answer CR items may not provide any unique information beyond a comparable MC item (Rodriguez, 2003), yet have much higher costs due to human scoring.

---

1 Rodriguez, M. C. (2002). Choosing an item format. In G. Tindal & T.M. Haladyna (Eds.), *Large-Scale Assessment Programs for All Students: Validity, Technical Adequacy, and Implementation* (pp. 213-231). Mahwah, NJ: Lawrence Erlbaum Associates.

Rodriguez, M. C. (2003). Construct equivalence of multiple choice and constructed response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40*(2), 163-184.

CR items are challenging to write so as to motivate intended responses without giving away the answer or confusing students – often such items are inadequately presented and fail to explicitly inform students about the expectations in a high-quality response (Rodriguez & Haladyna, 2013)2. Fewer items can be administered in the same time period as SR items, resulting in limited content coverage, and writing skill may interfere with the assessment of knowledge and skills in other domains.

The trade-offs are challenging. In particular, the challenge of writing CR items that go beyond the measurement potential of SR items is especially daunting. Rodriguez (2003) synthesized the empirical evidence on the equivalence of MC and CR items and illustrated that most CR items are written in such a way that they tap essentially the same knowledge, skills, and abilities as MC items. He argued that MC and CR items do in fact measure the same things when they are written to do so – that if the CR items really are intended to measure different aspects of the content domain or different cognitive skills, much more work needs to be done in the construction and scoring of CR items (for guidance on CR item development and scoring see Haladyna & Rodriguez, 2013; Lane and Swygert & Williamson, this volume).

…

Item Writing – A Collaborative Effort

…

A basic set of item specifications should include (Haladyna & Rodriguez, 2013):

1. Content domain and cognitive tasks to be included. In ECD, this includes the domain analysis of content and skills and specification of the domain model including claims and evidence.
   a. Description of the precise domains of knowledge and skills to be assessed, a guide to grade-level requirements;
   b. Guidance to support construct representation and comparability across tasks;
   c. Guidance for cognitive complexity;
   d. Intended targets for item difficulty;
   e. Standards and core elements of practice of professional societies and organizations.
2. Item formats allowed and the parameters around their structure. In ECD, this includes the assessment framework including the student, evidence, and task models.
   a. Sample or model items in each allowable format;
   b. Number of allowable options for each item format;
   c. Sources of and characteristics of reading passages;
   d. Sources and characteristics of stimulus materials (illustrations, figures, graphs);
   e. Sources and characteristics of practice-based cases, scenarios, vignettes;
   f. Issues related to diversity and local, regional, or national content relevance.
3. Item-writing guidelines to be followed.
4. Item editing style guide.
5. Process and criteria for item reviews.
6. Criteria for item selection.

2 Haladyna, T.M., & Rodriguez, M.C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Lifecycle of a Test Item

1. Test purpose, uses, and specifications are defined.
2. Item specifications are developed. Here we are assuming that the item specifications call for SR item formats. The decision to use SR items should be documented, presenting the argument supporting the appropriate and meaningful use of SR items to achieve the test's purpose.
3. Item writers are identified, selected, and trained, including a comprehensive introduction to steps #1 and #2 above, and training regarding item development for various subgroups including students with disabilities or English language learners. This may include the use of item generation techniques, such as the use of item shells or other models (see Gierl, this volume), including task models with the ECD approach.
4. Item writers engage in supervised item writing, iteratively writing and reviewing items with their peers, with the support of an item-writing leader.
5. Item writers continue in the process of item writing. Items are reviewed potentially by multiple groups:
    a. Peer item writers,
    b. Senior content specialists,
    c. Sensitivity review (for bias and fairness) including experts with relevant subgroups like persons with disabilities and English language learners,
    d. Measurement specialists,
    e. Copy editor.
6. Items are piloted or field tested, ideally as embedded items in operational tests. Items are then reviewed in several ways:
    a. Item analysis is conducted, including a review of the item difficulty and discrimination;
    b. Distractor analysis is conducted, to assess the functioning of the distractors (should be selected relatively uniformly and be selected more often by test takers scoring lower on the overall measure) (see Haladyna, this volume, for distractor analysis methods);
    c. Item analysis should include some form of DIF analysis, examining functioning across gender, race, language status (perhaps others as required by the testing authority);
    d. For new item types, consider conducting think-aloud cognitive interviews to establish (confirm) the cognitive task elicited by the item.
7. Decisions are made regarding the disposition of the item:
    a. Editing and revision,
    b. Elimination,
    c. Selection for inclusion in operational use.
8. Items selected for operational use are placed in the item bank, become available for operational tests, and are monitored for performance over time, until released or retired.

Gathering Validity Evidence to Support SR Item Development

Test items play important roles in the interpretative/use argument of achievement tests and thus in test validation. Kane (2013, and this volume) argued that as a test undergoes development, we also develop the interpretation/use argument, typically with a focus on identification of the kinds of evidence needed for the validity argument, including content-related evidence, generalizability analyses, studies of item functioning, think-aloud studies, and others. He suggested that potential challenges to the interpretation/use argument can be preempted through the collection of relevant data, during test development stages.

The Role of Items in the Interpretation/Use Argument

One important aspect of a common interpretation/use argument is the extrapolation inference, extending the interpretation from the universe of generalization to the target domain, making the "leap from claims about test performances to claims about the full range of performance in the target domain" (Kane, 2013, p. 28)3. Kane argued that the confidence we place in the extrapolation inference depends on the strength of the association between test design and the definition of the target domain – a function of item specifications and development procedures, item selection, and item scoring among others. Similarly, criterion-referenced score interpretations can enhance the interpretation argument by suggesting what test takers with various scores know and can do. Through a measurement model, like item-response theory, item analysis can inform us regarding the relation between ability and expected performance on items in the universe of generalization (Kane).

Haladyna and Rodriguez (2013) presented a list of potential inferences and assumptions regarding the interpretation/use argument for item development and thus a basis for item validation. They include in their list:

1. Organization of the target domain.
2. Organization of the universe of generalization.
3. Degree of fidelity between the universe of generalization and target domain.
4. Selection of item formats to achieve the test purpose and intended inferences.
5. Item development process.
6. Item content and cognitive demand.
7. Item review process and results.
8. Item pretest evidence and selection criteria.
9. Item contribution to the internal structure of the test.
10. Process of item revision, selection, elimination.

Such information can be collected and documented in a technical manual (see Ferrara & Lai, this volume), an essential report documenting the process and outcomes of test development, and within that, item development. As the interpretation/use argument is articulated and the intended inferences and assumptions are uncovered, the validity argument can be strengthened, documented, and reported in an effort to enhance the interpretation and use of tests.

---

3 Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Excerpts from:

Rodriguez, M.C., & Haladyna, T.M. (2013). Objective testing of educational achievement. In K. Geisinger (Ed.), *APA Handbook of testing and assessment in psychology* (pp. 305-314). Washington DC: American Psychological Association.

Selected-Response Item Formats

Conventional MC

When it comes to describing the distribution, the standard deviation tells us

A. where most of the scores are located.
B. if the distribution is normal.
C. how far the scores are spread out.

Alternate-Choice

If a distribution of raw scores is positively skewed, converting to *T*-scores will result in what type of distribution?

A. Normal
B. Positively skewed

True-False (Dichotomous-Choice)

True or False: If the item difficulty index is .70, then 30% of examinees answered the question correctly.

Multiple True-False

Consider the following actions that may affect test score validity evidence. Determine whether each is True or False.

1. Adding more test items of similar quality improves test score validity.
2. Increasing the sample size will increase criterion-related validity correlations.
3. Obtaining a sample with more test score variability increases criterion-related validity correlations.
4. Eliminating items with poor item-total correlations (discrimination) will improve content-related validity evidence.

Matching

Match each term on the right with the description on the left.

1. score stability _____
2. attention deficit disorder _____
3. content alignment _____
4. *p*-value _____
5. item-total correlation _____

A. systematic error
B. random error
C. item difficulty
D. item discrimination
E. reliability evidence
F. validity evidence

Context-Dependent Item Set
An anonymous Standard Item Analysis Report was found online that appears to be a cumulative report for an exam in a specific course. This exam has been completed by 327 students. The total number of items is 50. Refer to this Item Analysis Report [not shown in this example] when answering the following items.

1. Which item is the easiest? _____
2. Which item should be revised into a true-false item? _____
3. Which item has the best discrimination? _____
4. Identify <u>one</u> item that has the best example of effective distractors? _____
5. Identify <u>one</u> item that is most likely to have two correct answers? _____

Complex Multiple-Choice (use of this type is not recommended, see below)
Which are norm-referenced interpretations of test scores?

1. John's score is three standard deviations above the class mean.
2. Mary answered 80 percent of the items correctly.
3. Eighty percent of the class scored above a T-score of 45.
4. The average math score for Arlington High is equal to the district average.
5. Antonio is proficient in 5^{th} grade reading.

A. 1 and 3.
B. 2, 3 and 5.
C. 2 and 5.
D. 1, 3 and 4.
E. All 5.

Excerpts from:

Haladyna, T.M., & Rodriguez, M.C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Table 6.1
Guidelines for Writing SR Items

CONTENT CONCERNS

1. Base each item on one type of content and cognitive demand.
2. Use new material to elicit higher level thinking.
3. Keep the content of items independent of one another.
4. Test important content. Avoid overly specific and overly general content.
5. Avoid opinions unless qualified.
6. Avoid trick items.

FORMATTING CONCERNS

7. Format each item vertically instead of horizontally.

STYLE CONCERNS

8. Edit and proof items.
9. Keep linguistic complexity appropriate to the group being tested.
10. Minimize the amount of reading in each item. Avoid window dressing.

WRITING THE STEM

11. State the central idea clearly and concisely in the stem and not in the options.
12. Word the stem positively, avoid negative phrasing.

WRITING THE OPTIONS

13. Use only options that are plausible and discriminating. Three options are usually sufficient.
14. Make sure that only one of these options is the right answer.
15. Vary the location of the right answer according to the number of options
16. Place options in logical or numerical order.
17. Keep options independent; options should not be overlapping.
18. Avoid using the options *none-of-the-above, all-of-the-above,* and *I don=t know*.
19. Word the options positively; avoid negative words such as NOT.
20. Avoid giving clues to the right answer:
    a. Keep the length of options about equal.
    b. Avoid specific determiners including always, never, completely, and absolutely.
    c. Avoid clang associations, options identical to or resembling words in the stem.
    d. Avoid pairs or triplets of options that clue the test taker to the correct choice.
    e. Avoid blatantly absurd, ridiculous options.
    f. Keep options homogeneous in content and grammatical structure.
21. Make all distractors plausible. Use typical errors of test takers to write distractors.
22. Avoid the use of humor.

Table 11.1
Guidelines for Writing CR Items

CONTENT CONCERNS

1. Clarify the domain of knowledge and skills to be tested.
2. Ensure that the format is appropriate for the intended cognitive demand.
3. Ensure construct comparability across tasks.

FORMATTING & STYLE CONCERNS

4. Edit and proof instructions, items, and item formatting.
5. Pilot items and test procedures.

WRITING THE DIRECTIONS/STIMULUS

6. Clearly define directions, expectations for response format, and task demands.
7. Provide information about scoring criteria.
8. Avoid requiring implicit assumptions; avoid construct-irrelevant task features.

CONTEXT CONCERNS

9. Consider cultural and regional diversity and accessibility.
10. Ensure that the linguistic complexity is suitable for intended population of test takers.

Table 12.1
CR Scoring Guidelines

CONTENT CONCERNS

1. Clarify the intended content and cognitive demand of the task as targets for scoring.
2. Specify factors in scoring that are irrelevant to the task demands.

SCORING GUIDE DEVELOPMENT

3. Select an appropriate scoring method.
4. Begin scoring guide development during task construction (Item writing).
    a. Clarify distinctions across score points.
    b. Define clear justifications within score points.
    c. Do not over specify expected responses.
    d. Expectations for the same cognitive demand should be the same across similar tasks and scoring rules.
5. Review actual responses to refine scoring guide.

SCORING PROCESS

6. Qualify raters.
7. Train raters.
8. Rate consistently.
9. Minimize bias.
10. Obtain multiple ratings.
11. Monitor Ratings.

The Fairness Review

As stated by Zieky4, fairness review is intended to detect any construct-irrelevant factor that may affect test scores and item responses. The intent is to increase validity, not to promote censorship or political correctness. Zieky provides six guidelines that should drive this review:

1. *Treat people with respect.* This includes employing a diverse set of proper names for the characters in items; they should represent the population of test takers. Ethnocentrism should be avoided. Test items should acknowledge the diversity inherent in each subgroup. By achieving each element of fairness presented here, the item writer makes great strides in treating test takers with respect.
2. *Reduce the effects of construct-irrelevant variance (CIV).* Haladyna and Downing5 provide an extensive treatment of this topic. CIV is systematic error. Any factor that increases or decreases the difficulty of the item for a subgroup of those being tested produces CIV. Such influences should be removed or the item should be retired. These untoward influences might include irrelevant or difficult-to-understand charts or graphs, difficult or inappropriate vocabulary, specialized vocabulary, technical language, inappropriate or undocumented acronyms, religious or cultural references, and the like. The ETS guidelines (ETS, 2003) provide more specific instances.
3. *Avoid emotionally charged content*. This advice is based on subjective judgments and requires the use of SMEs with sensitivity to potentially upsetting material. Some offensive topics to avoid include abortion, cultural taboos, genocide, Halloween, murder, political view, rape, sexual orientation, supernatural events, suicide, religion, and torture.
4. *Use appropriate terminology*. This fairness review category is aimed at distinctions we draw in describing people: African-American, Black, Negro, or Colored? Native American or Indian? In most instance, such designations for people are irrelevant. However, if these designations are used, they should be done with respect to the group of people being identified. Maintaining parallelism is also important. For example, characters in test item vignettes should be of equal status when it comes to gender, race, and ethnic background. General Mills in the U. S. Army can be a man or a woman. However, it would not be proper to use terms such as a *male nurse* or *woman scientist*. Male-dominated terms such as *man-made* are unacceptable.
5. *Avoid stereotypes*. Although we often forget, we should not assume that a character in a vignette is of one gender category or one ethnic classification. For example, all people in child care are women, all teachers are women, all firemen are men, all dental hygienists are women. As stated above, test items, reading passages, and associated materials, should acknowledge the diversity within subgroups.
6. *Represent diversity*. Historically, we have depicted characters in test items as coming from one race or social class in an innocent, idealistic way. Fairness in testing suggests that more diversity should be represented. Characters in test items should come from many social climes and ethnic backgrounds. If a person's gender, race, education, or social standing is relevant, the reference should be proportional with the population and fair.

---

4 Zieky, M. (2006). Fairness reviews in assessment. In S. M. Downing & T. M. Haladyna (Eds.) *Handbook of test development* (pp. 359-376). Mahwah, NJ: Lawrence Erlbaum Associates.

5 Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Table 16.4
Linguistic Modifications of Test Items

Feature	Brief Description
Word frequency/familiarity	Test items should use words that have a high frequency in common literature for that group of test takers.
Word length	Longer words tend to be less familiar and should be avoided where possible.
Sentence length	Not only should sentences be shorter but their grammatical structure should be simple.
Voice of Verb Phrase	Passive voice should be avoided.
Length of Nominals	Noun phrases with several modifiers are troublesome to ELL test takers.
Complex question phrases	Longer questions with a complex structure are also troublesome to ELL test takers.
Comparative structures	Comparative constructions are another source of construct-irrelevant difficulty to test takers.
Prepositional phrases	ELL students have difficulty with prepositional phrases.
Sentence and discourse structures	Sentences in a paragraph may have different discourse structures that confuse some ELLs.
Subordinate clauses	Subordinate clauses are more complex than coordinate clauses and thus challenge the ELL.
Conditional clauses	Conditional and adverbial clauses contribute to CIV difficulty.
Relative clauses	Some test takers have limited exposure to relative clauses, which may cause them to perform lower than expected.
Concrete versus abstract presentations.	Narrative presentations than to be better understood than expository presentations.
Negation	Negation is harder to comprehend. Negation is not recommended as a general item-writing principle.

Source: Abedi, 2006.

Abedi, J. (2006). Language issues in item development. In S. M. Downing & T. M. Haladyna (Eds.) *Handbook of test development* (pp. 377–398). Mahwah, NJ: Lawrence Erlbaum Associates.