

The Origin of Selected-Response Items

Michael C. Rodriguez

Educational Psychology
University of Minnesota

June 2014

The Origin of Selected-Response Items

The origin of the very first selected-response (SR) item may never be known for certain, but early examples of these item types can be found in the literature nearly 100 years ago. We have some evidence of the earliest forms of assessment taking place in China 4000 years ago in the form of civil-service performance assessments and modern records of oral and written exams as early as the 1200s at the University of Bologna, 1400s at Louvain University, and during the 1500s in Jesuit schools (DuBois, 1970). By the mid-1800s, short-answer items were appearing in school district assessments in Boston and it appears that Michigan developed the first state testing program in the late 1800s (Ed Roeber, personal communication, March 28, 2013); Michigan is also home to the nation's first state department of education.

Although it is possible that these early district and state tests may have had some forms of SR items, DuBois (1970) attributed the origin of the multiple-choice (MC) item to the Army Alpha in 1917. The Army Alpha was the first large-scale standardized assessment with MC items as we know them today, but not the first publically available test to employ them. Once the College Board began using the MC item type in college entrance exams, its use quickly gained popularity. The first formal and systematic treatment of the item type was delivered by Ebel (1951) in his landmark chapter on item writing in the first edition of *Educational Measurement*.

In attempt to identify the earliest instance of the use of SR item types, following the lead of Haladyna and Rodriguez (2013) and their reference to a SR test of silent reading by Monroe (1918), I uncovered an article reporting on the measurement of achievement in English grammar by Starch (1915). A brief review of his work will be instructive.

In his measure, Starch (1915) tested grammar and punctuation. As examples of items on the grammatical correctness measure, consider the following instructions and items:

Each of the following sentences gives in parenthesis two ways in which it may be stated. Cross out the one that you think is incorrect or bad. If you think both are incorrect cross both out. If you think both are correct underline both. (p. 615)

1. It was so misty we (could hardly; couldn't hardly) see.
 2. The gazette reported (he; him) to be dead.
 3. He was a patriot, but all the rest (were traitors; traitors).
-

Today, we would call this format alternate choice. Regarding the options, one, both, or neither could be correct, where the distractors are plausible, conveying common errors in grammar. Starch argued that the advantages of the scale included definiteness, objectivity, comparability, and convenience, all attributes we would attach to modern SR tests. He interestingly added: "With the help of a key, any person, who may know nothing about grammar or correct usage, can score the tests as well as a linguist can" (p. 619), which must have been considered a great reduction in burden at a point in time when written exams (performance assessments) were state of the art. Today we enjoy automated scoring and avoid the linguist, or human scorer, altogether!

Another form of an SR item test was also offered in this 1915 paper by Starch, in the English grammar test of parts of speech. A 200 word narrative selection was given, for which the test taker was to indicate the correct part of speech by placing above each word in the text, the abbreviation for its part of speech: noun (n), pronoun (pro), adjective (adj), verb (v), adverb (adv), preposition (pre), conjunction (c), and interjection (i). This resulted in 200 responses (one per word in the provided text) with eight options from which to select a correct response. This might be considered a MC item with a common set of eight options, or even a matching exercise with a list of 200 reference words and eight option choices that could be used multiple times (or not at all).

Other similar measures were offered by Starch in this early article on the development and use of SR tests (although he never referred to them as selected response).

Based on the work of Starch, Kelly (1916) offered an early iteration of the *Kansas Silent Reading Tests*. These tests also contained SR formatted items, such as the following:

-
4. Think of the thickness of the peelings of apples and oranges. Put a line around the name of the fruit having the thinner peeling.

apples oranges

5. Three words are given below. One of them has been left out of this sentence: I cannot _____ the girl who has the flag. Draw a line around the word which is needed in the above sentence

red see come

6. In the following words, find one letter which is contained in only three of them and then cross out the word which does not contain that letter.

ail thief live anvil

It is interesting to note that the number of options varied among the items, with some including up to six options. In 1917, Monroe reported on the use of the *Kansas Silent Reading Tests* and discussed a number of criticisms that had been made of the test at that time. The most frequent criticism was one regarding validity:

the exercises partake of the nature of puzzles, a very large number of them being arithmetical in nature. For this reason, they are not typical reading material. ...due largely to the fact that an effort was made to have the tests [be] objective; that is, to have them of such a nature that a minimum of opinion would be exercised in marking the papers. This is a feature which is very important and which is not true of most of the reading tests which have been devised. The question might well be raised whether or not it was worthwhile to sacrifice other things to obtain this feature. (p. 608)

Without spending too much time on reviewing these early efforts, I will close with the subsequent attempt of Monroe (1918) to develop a revised form of the *Kansas Silent Reading Tests* that focused more on the skills of reading (and of course the debate regarding what constitutes reading skills continues today).

The questions remained similarly structured:

-
7. The dog lay down. The rooster flew to the top of a tree and the cat climbed to one of the branches. Before they went to sleep the rooster saw a light in the forest. He called to his friends. Where was the light the rooster saw?

sky house barn wagon forest

8. They rested and talked. Their talk was all about their flocks, a dull theme to the world, yet a theme which was all the world to them. What do you suppose was the occupation of these men?

carpenter doctor merchant shepherd blacksmith

In these early attempts to explore SR items for educational measurement, we find many of the issues that remain today regarding the development and validation of SR items, including format selection, specifying the construct for item development, validation of the target of measurement, number of options, distractor relevance or plausibility, and others.

References

- DuBois, P.H. (1970). *A history of psychological testing*. Boston, MA: Allyn & Bacon.
- Ebel, R.L. (1951). Writing the test item. In E. F. Lindquist (Ed.), *Educational Measurement* (1st ed., pp. 185-249). Washington, DC: American Council on Education.
- Haladyna, T.M., & Rodriguez, M.C. (2013). *Developing and validating test items*. New York, NY: Routledge.
- Kelly, F.J. (1916). The Kansas Silent Reading Tests. *The Journal of Educational Psychology*, 7, 63-80.
- Monroe, W.S. (1917). A report on the use of the Kansas Silent Reading Tests with over one hundred thousand children. *The Journal of Educational Psychology* 8, 600-608.
- Monroe, W.S. (1918). Monroe's Standardized Silent Reading Tests. *The Journal of Educational Psychology*, 9, 303-312.
- Starch, D. (1915). The measurement of achievement in English grammar. *The Journal of Educational Psychology*, 6, 615-626.