Guidelines and Suggestions for Avoiding Cultural Bias in Multiple-Choice Test Questions

Joel P. Wiesen, Ph.D., Director
Applied Personnel Research
27 Judith Road
Newton, Massachusetts 02459-1715
wiesen@personnelselection.com

Conference: 17th Annual Conference of the Society for Industrial and Organizational Psychology; Toronto, Canada, 4/14/2002
Session Title: Developing Defensible Written Test Questions: Art, Science, and Some Guidelines
Presentation Title: Review of Written, Multiple-Choice Test Items, with Some Emphasis on Cultural Bias

1. Readability of Test Questions

If reading is not a pervasive aspect of all parts of the job, keep the reading level of most of the test below that needed on the job; otherwise your measure of various (non-reading) abilities will be confounded with (i.e., reflect) reading ability. For most tests, only the reading questions should be at the reading level needed on the job.

Use readability formulas carefully. They are measurement tools, not guides to clear writing. Questions may be difficult to read despite low readability measures if you use hard words (including short hard words, such as acute), passive tense, or complex sentences. It is best to use clear, plain English.

Manually check the readability indices found in major word processors to be sure they are working as intended. For example, one major word processor takes every period to indicate the end of a sentence. So if the questions are numbered and each question number is followed by a period, each question number is counted as a sentence. That makes for artificially low readability statistics.

There should be no (unintended) errors in grammar or spelling.

2. Reading Lists

Often reading lists are a good idea because they allow people the opportunity to prepare for the test. To allow for sufficient time to prepare, a reading list should best be published long before the examination date. Six months lead time is the most I've recommended. All people should have an equal length of calendar time to study.

3. Unnecessarily Academic Test Content

When textbooks are used as source material for jobs without an academic requirement, there is a serious risk that they introduce problems akin to those discussed above concerning readability. Often textbooks are written for a college audience, and the reading level may be too high, and the content overly academic for the job in question. Some texts are much better and others much worse in this regard. This should be a primary consideration in choosing texts to appear on a reading list.

4. Everyday, Practical Reference Material

Although authors of current college textbooks tend to try to make their textbooks readable, the content and presentation found in many books is quite academic, with much jargon and picayune material. Sometimes it is possible to base test questions on local SOPs, ISO9000 documentation, local rules or laws or guidelines. These types of source material usually are more focused on practical aspects of the job than are textbooks. The use of local rules, regulations, operating procedures, and guidelines is generally desirable to the extent that these are current and reflect what is or should be done on the job.

5. Test Content Which Unnecessarily Evokes Emotional Responses

Some questions may arouse anxiety or other emotional response from some candidates. They may cause unequal testing conditions. For example, it would be best to avoid a table depicting incarceration rates by ethnic and religious group. Avoid sensitive topics (e.g., criminal justice system, HIV/Aids) unless they are germane to the subject of the test. Some words or phrases have more than one meaning, and one of these meanings might have emotional loading for only certain persons or groups; for example, courier (letter vs drug), court (romance vs criminal justice). Other words or phrases reflect or convey bias or exclusion; for example, fair-haired child or cross your fingers. Such words and phrases might best be avoided unless they are germane to the subject of the test. Avoid presenting age, gender, culture, or ethnic groups in a stereotyped fashion, even in wrong answers.

6. Balanced or Neutral Representation as to Age, Gender, Culture, and Ethnic Group

One approach to this would have a balance of male and female names mentioned, rather than just one gender. If some people mentioned in the questions are in supervisory roles, they should not all be one gender or ethnic group. Likewise for criminal roles, leadership roles, etc. One should strive for test content which puts all test takers at ease. Avoid presenting age, gender, culture, or ethnic groups as homogenous in nature.

7. Contamination With Irrelevant Test Content

Sometimes questions which are intended to measure one area are also measuring another. Reading level (discussed above) may be the most common example of this, but there are others. For example, math items which ask about import duties or increases or decreases in stock prices may confuse people who do not travel internationally or do not trade in stocks. Verbal analogy questions which use relatively obscure words will confound the measurement of reasoning ability with work knowledge. (It is easy to fall into this trap, perhaps because many common words have more than one meaning, and many more precise words are relatively obscure.)

8. Test Content Which is Equally Familiar to All Groups

Most test areas can be measured with a questions in a variety of sub-areas. Often this is not mentioned in the test outline. Sometimes one or more areas are less familiar to some groups of candidates. For example, math questions which involve import duties or tariffs may be less familiar to people from lower socioeconomic groups. As another example, mechanical aptitude questions which focus on the working or repair of cars may be less familiar to women, and questions which focus on the working or repair of sewing machines may be less familiar to men. However, there may be questions about common household objects which would test mechanical aptitude and be equally familiar to men and women.

The wording of test questions should also be equally accessible to all groups. Avoid slang and colloquial wording, and any wording which may be differentially familiar to various groups. Strive for equal familiarity of words for all groups (both connotations and denotations).

All groups should have an equal familiarity with the test format or be given sufficient time and experience to ameliorate differences in such familiarity.

9. Context

The context of the test questions or instructions should not be confusing. For example, if the instructions for a verbal analogy test include very dated words or concepts (e.g., inkwell and blotter) then test takers might think that every question should be evaluated in terms of possible dated meanings. This might have a greater negative effect on the less test wise candidates. Sometimes questions presume a context which is not stated, making the question tricky or just faulty (e.g., asking a question about an interview on a test for police sergeant and not specifying that it is in the context of an employee performance evaluation as opposed to an interview of a witness). Sufficient context should be given in the stem (e.g., rather than ask "When selecting a landing area you should..." it would be clearer to ask, "When selecting a landing area for an LX2 helicopter you should...") Asking about a non-standard application of a familiar principle can be tricky. Even the numbering of the questions and the answer sheet can be confusing and thereby contribute to confusion and errors in filling in answer bubbles (e.g., when the numbering goes across the page rather than down the side, or when there are many bubbles on the answer sheet which are printed far from their labels.) In printing the examination, skipping lines between questions can help avoid confusion. Quality of reproduction can be so poor as to make questions difficult to read.

10. Tricky Questions (Some of this section is based on Roberts, D.M.., 1993.)

Tricky questions should be avoided. Questions can be tricky for reasons of content, context, or wording. With respect to content, questions may be tricky due to: trivial nature of the key or the question, voluminous extraneous material in the question which goes beyond what is found on the job, multiple correct answers, or very fine distinctions between distracters and the key.

With respect to context, a hard question mixed in with easy questions may be tricky because test takers get used to answering based on easy considerations and do not consider subtle issues. (Also, see previous section on context.)

With respect to wording, questions may be tricky due to: unclear wording of question, use of common words but with an unusual meaning, use of unusual words when usual words are commonly used, or use of double negatives. With some applicant groups, the use of negatives can be tricky, and it can be helpful to capitalize words like not, and least.

11. Time Limits

Highly speeded tests should be undertaken with caution. Beyond trying to minimize the discomfort of candidates, there is some evidence that blacks are at a relative disadvantage in taking speeded tests (Hartigan and Wigdor, 1989, page 108).

12. Equal Practice in Test Taking

Exposure to sample questions no doubt reduces anxiety. Beyond that, one study showed that people retested on an alternate version yielded a mean increase of .2 SDs for cognitive tests (Hartigan and Wigdor, 1989, page 112).

13. Equal Access to Test Coaching

Inequity in access to commercial coaching may result in test bias (perhaps only when evaluated against an external criterion, and so missed with the more often used internal measures of bias).

References

Hartigan, John A. and Wigdor, Alexandra K. (1989) *Fairness in Employment Testing*. Washington, D.C.: National Academy Press.

Roberts, D.M. (1993) An empirical study of the nature of trick test questions. *Journal of Educational Measurement, 30,* 331-334.