

SWEDISH

How to develop a good test and good test questions

Susan Richardson, Ph.D., DABR, FAAPM

SWEDISH
CANCER INSTITUTE

SWEDISH

COI/Disclosure

- I have no relevant disclosures
- I examine and write questions for the ABR but I am not here in that capacity

SWEDISH
CANCER INSTITUTE

SWEDISH

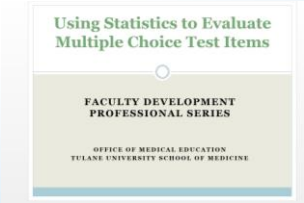
Layout

- What makes a good test?
- How do I build my test and test questions?
 - What should I experience as a test taker (SAMS, ABRs, etc)

SWEDISH
CANCER INSTITUTE

SWEDISH

Real science for scientists




SWEDISH CANCER INSTITUTE

SWEDISH

What is the purpose of testing?

- To communicate what you view is important
- Motivate students to learn/study
- To assess skill in interpreting data and information
- To assess skill in decision making

These combined help build important components of clinical skills.




SWEDISH CANCER INSTITUTE

SWEDISH

Build a test in 10 easy steps...

1. Conducting the Job Task Analysis
2. Developing the Test Blueprint
3. Developing Items
4. Reviewing and Validating Items
5. Assembling and Delivering Beta Exams
6. Analyzing Beta Exam Results
7. Constructing Equivalent Exam Forms
8. Establishing the Passing Score
9. Administering/Scoring Operational Exams
10. Providing Ongoing Test Maintenance

www2.sas.com/proceedings/sugi25/25/pa/25p244.pdf



SWEDISH CANCER INSTITUTE


SWEDISH

Developing a test

- The test should be **VALID**
 - It should test what you think it tests or what it claims to test.
- The test should be **RELIABLE**
- The test should be **FAIR**

• A reliable test may not be valid! And vice versa.

SWEDISH CANCER INSTITUTE




SWEDISH

Validity

- There are many different types of validity. Too many to actually discuss in one hour.
- Ideas to increase validity:
 - Have experts independently review test items
 - Clear, simple instructions
 - Evaluate the correlation of the curriculum you are testing and the number of test questions associated with the components.
 - The number of items of each topic should be related to the time and criticality in the course.

<http://files.eric.ed.gov/fulltext/ED501716.pdf>

SWEDISH CANCER INSTITUTE




SWEDISH

Reliability

- Reliability, which is the best single measure of test accuracy, is the extent to which test results are consistent, stable, and free of error variance.
- It is the extent to which a test provides the same ranking of examinees when it is re-administered and is measured by Coefficient Alpha or KR-20.
- There are many different types of reliability coefficients such as stability, equivalence, and internal consistency.

SWEDISH CANCER INSTITUTE



SWEDISH

Kuder-Richardson Formula 20

From Wikipedia, the free encyclopedia

In psychometrics, the Kuder-Richardson Formula 20 (KR-20) first published in 1927^[1] is a measure of internal consistency reliability for measures with dichotomous choices. It is analogous to Cronbach's α , except Cronbach's is also used for non-dichotomous (continuous) measures.^[2] It is often claimed that a high KR-20 coefficient (e.g., $\rightarrow 0.95$) indicates a homogeneous test. However, like Cronbach's α , homogeneity (that is, unidimensionality) is actually an assumption, not a conclusion, of reliability coefficients. It is possible, for example, to have a high KR-20 with a multidimensional scale, especially with a large number of items.

Values can range from 0.00 to 1.00 (sometimes expressed as 0 to 100), with high values indicating that the examination is likely to correlate with alternate forms (a desirable characteristic). The KR-20 may be affected by difficulty of the test, the spread in scores and the length of the examination.

In the case where scores are not tau-equivalent (for example when there is not homogeneity but rather examination items of increasing difficulty) then the KR-20 is an indication of the lower bound of internal consistency (reliability).

The formula for KR-20 for a test with K test items numbered 1 to K is:

$$r = \frac{K}{K-1} \left[1 - \frac{\sum_{j=1}^K p_j^2}{p^2} \right]$$

where p_j is the proportion of correct responses to test item j , p is the proportion of correct responses to test items (i.e. that $p_j = p$, $j = 1, \dots, K$), and the variance of the denominator is:

$$p_j^2 = \frac{\sum_{j=1}^K (X_j - \bar{X})^2}{n}$$

where n is the total sample size.

If it is important to use unbiased operators then the sum of squares should be divided by degrees of freedom ($n - 1$) and the probabilities are multiplied by $\frac{n}{n-1}$.

Since Cronbach's α was published in 1951, there has been no known advantage to KR-20 over Cronbach's α . KR-20 is seen as a derivative of the Cronbach formula, with the advantage to Cronbach that it

SWEDISH
CANCER INSTITUTE

SWEDISH

Fairness

- Fairness is a social rather than a psychometric concept.
 - Equal group outcomes? (typically a rejected definition)
 - Equitable treatment of all examinees?
 - Lack of predictive bias?

<https://www.siop.org/>

SWEDISH
CANCER INSTITUTE

SWEDISH

How do I write a good test question?

- Two components:
 1. It must address important content
 2. It must be well-structured


SWEDISH
CANCER INSTITUTE

SWEDISH

Constructing the Question

- Three parts:
 1. Item = The test question.
 2. Stem = question itself. Contains background information along with the request for an answer.
 3. Options = Key and distractors = answers

SWEDISH CANCER INSTITUTE




SWEDISH

Written exams

- Typically two types of questions
 1. True/False
 - Difficult to write and interpret because the test taker must examine each option and determine HOW true or HOW false each answer is
 - Often makes the test taker guess at what the item writer or grader had in mind
 2. Best answer or multiple choice

SWEDISH CANCER INSTITUTE




SWEDISH

True and False Questions

- Typically involve lots of guessing
- Actually hard to write questions that are unequivocally true or unequivocally false
- Tests trivial knowledge
- Not very discriminatory
- (are often very reliable!)
- Abraham Lincoln was born in a log cabin in Springfield, Missouri. True or False?

SWEDISH CANCER INSTITUTE



True and False

- Stems must be clear and unambiguous.
- Phrases that should be avoided
 - is useful for
 - is important
 - Could be
 - May be
 - usually
- Words that are vague or requiring cueing should be avoided
- Options must be absolutely true or false
 - No 50 shades of gray allowed.



A word on True/False

- The general recommendation by the NBME is that true/false questions are not to be used.
 - While easier to write, they are more problematic
 - They usually require the test taker to recall an isolated fact (something we try to avoid)
- All ABR/MOC/SAM questions are in the multiple choice format.




What's in a Multiple Choice question?

- STEM
 - This is the "question"
 - Contains background and situational information
 - Should end in "?"
 - Should be written in present tense
- OPTIONS
 - These are the choices that the person can make
 - One is the KEY (the right answer)
 - The rest are DISTRACTORS

Least Correct

Most Correct






Example of truth scales


The way to a man's heart is through his

1. aorta
2. pulmonary arteries
3. pulmonary veins
4. stomach

<i>Totally Wrong Options</i>	<i>Totally Correct Options</i>
------------------------------	--------------------------------


Example courtesy of nbme.org






The STEM

- The stem should be **FOCUSED**
 - What area of knowledge am I trying to test?
 - Test only a single concept with each question
 - The concept should be important and relevant.
 - "which of the following is correct?" is unfocused.
- The stem should be **CLEAR**
 - You shouldn't try and trick the test taker
 - Lose any extraneous or ambiguous information






More about the STEM

- The stem should be **CONCISE**
- The stem should assess knowledge, not fact recall

<ul style="list-style-type: none"> • The PDD for a 6MV linac at 10 cm is: <ol style="list-style-type: none"> A. 97% B. 87% C. 77% D. 67% 	<ul style="list-style-type: none"> • The PDD for a 6MV linac at 10 cm is: <ol style="list-style-type: none"> A. Greater than an 18 MV beam B. Less than an 18 MV beam C. The same as an 18 MV beam
--	---



SWEDISH
CANCER INSTITUTE

More about the STEM

- A good question is **POSITIVELY** worded
 - "Which of the following" NOT "Which of the following does not"
 - "All of the following EXCEPT"
- The order should be in a logical and linear format:
 - Background information + situational data + request for answer
 - Eg: A 6MV linear accelerator is being calibrated for absolute dose, which is the most appropriate detector for this measurement?

SWEDISH
CANCER INSTITUTE

SWEDISH
CANCER INSTITUTE

The COVER test

- The STEM needs to be a question.
- The "Cover" test means – can you more or less answer the question without seeing the options?
 - Example: Ion chambers:
 - Are absolute dosimeters
 - Are filled with noble gases
 - Are constructed of stainless steel
- Better example: An example of an absolute dosimeter is which of the following?

SWEDISH
CANCER INSTITUTE

SWEDISH
CANCER INSTITUTE

The KEY & Distractors

- Most question writers spend the most time focusing on the key when all options should have equal time and effort spent on them
- Terminology:
 - Avoid using the terms "always" and "never"
 - Other words to avoid: usually, commonly, sometimes, frequently, often, rarely, most of the time, etc.
 - While these seem to be contradictory, they both have implications which rely on a judgment call and an interpretation by the test taker.

SWEDISH
CANCER INSTITUTE

The KEY and distractors, cont.

- The answers should be uniform in nature (e.g. all numerical)
- The distractors should be short and not contain extra information
- A subset of the distractors should not be collectively exhaustive nor overlap:

Example: The best physicists are:

- A) People over 40
- B) People under 40
- C) People who are 40
- D) Generation X-ers

The KEY and distractors, cont.

- The key and the distractors should be equal in length:

Example: Which of the following describes the radial dose function in the TG43 formalism?

- A) $G(r, \theta)$
- B) S_k
- C) Λ
- D) The factor that accounts for dose fall-off on the transverse plane due to photon scattering and attenuation that excludes the fall off that is included by the geometry function and is equal to unity at r_0 at 1 cm.

The KEY and distractors, cont.

- None of the above is not allowed
- All of the above is not allowed
 - These are essentially true/false questions in disguise.
- The KEY (and distractors) should not contain exact words or forms of the word in the question (word repeats)

E.g. What is the appropriate dosimeter for measuring ionization?

- Calorimeter
- Ionization chamber
- Film
- Diode

Matched pairs in the distractors

- It is ok to 'match' options in the distractors as long as they include all treatment options (one pair is not ok).

Example: In men with prostate-specific antigen (PSA) levels between 4.1 and 10.0 ng/mL, what effect do lower percent-free PSA levels have?

- A) They increase the likelihood of cancer diagnosis.
- B) They decrease the likelihood of cancer diagnosis.
- C) They increase the likelihood of cancer survival.
- D) They decrease the likelihood of cancer survival.



Item difficulty and distractors

- It is easy to alter the difficulty of the question by altering the distractors

Who is giving the lecture I'm listening to right now?

- A. Some person named Susan
- B. Some person name Chuck
- C. Mickey Mouse
- D. Donald Trump

vs

Who is giving the lecture I'm listening to right now?

- A. Some person named Susan
- B. Some person named Sarah
- C. Some person named Sierra
- D. Some person named Sandy



Distractors

- Distractors need to be clearly incorrect (not just not the best answer)
- But they must also be plausible or realistic
- If no examinees choose that distractor because it is so unrealistic, it doesn't help the exam (or you) at all



Summary and Conclusions

- Creating a good exam is:
 - Easy
 - Hard
 - A Lot of work
 - Time consuming
 - Scientific and Rule based
 - All of the above?



Resources

- National Board of Medical Examiners publication:
<http://www.nbme.org/publications/item-writing-manual.html>