Writing Items For Psychological Tests

Merle Canfield

Criterion or Outcome Measure:

Two synonyms for this concept are outcome measure and criterion variable. It is the measure of the amount of some mental health. It could be a single question, a questionnaire, a multi dimensional questionnaire or an extensive analysis.

Treatment Measure:

This may also be referred to as the treatment variable. This is the amount of a specific treatment administered to a client.

Converting psychological objects (goals, attitudes, feelings, traits or behavior) into observable situations is termed psychometry. That is, to make statements so that they can be used to determine to what degree a psychological object is present. Such statements will now be referred to as items and tests (a test being one or more items).

Items are designed to will indicate some quantity of the psychological object. There are two ways which this can be done: The first is that quantification be contained in the statement itself. The second is that the statement contain only the psychological object and quantifying statement or numbers follow the stem. Which of these methods you choose will depend upon your needs, the nature of your statements and the validity in the various methods.

The following are examples of four popular methods which includes instructions and three sample items:

Method #1 (Likert Type)

This scale has been prepared so that you indicate how you feel. Please respond to every item. In each case, draw a circle around the letter which represents your own reaction as follows:

SA if you strongly agree with the statement
A if you agree but not as strongly
N if you are neutral
D if you disagree but not too strongly
SD if you strongly disagree

Remember the only correct answer is the one which actually represents how you feel.

1. I am nervous. SA A N D SD

2. I get nervous easily. SA A N D SD

3. Not many things bother me. SA A N D SD

Method #2 (Modified Likert)

Instructions would be the same as in Method #1. Items would change as follows:

1.	I am nervous	Never	Infrequently	Now and Then	Often	Always
2.	I get nervous easily	Never	Infrequently	Now and Then	Often	Always
3.	Things bother me.	Never	Infrequently	Now and Then	Often	Always

Method #3 (Thurstone)

Below is a series of statements about your feelings. Read each statement and put a check beside the statement that reflects how you feel.

1. I sometimes get nervous. ____

2. Lots of things bother me. ____

3. Nothing bothers me. ____

Method #14 (Semantic Differential)

This scale has been prepared so that you can indicate how you feel. There are pairs of words with opposite meanings. Place a check in the blank closet to the word that indicates your feelings. If you are neutral, place a check in the middle blank.

1. Anxious ____ ____ ____ ____ ____ ____ ____ Calm

2. Excited ____ ____ ____ ____ ____ ____ ____ Indifferent

3. Elated ____ ____ ____ ____ ____ ____ ____ Depressed

Method I (Likert) could also be changed to true or false. That is, either it did occur or it did not.

The Likert method is recommended as a standard; the other methods are for special cases. The problem with true-false formats (and consequently the check-list) is that it produces dichotomous data and in a sense is a forced-choice item. Statistical power is lost in dichotomous data, and some participants indicate that they cannot respond with their true feelings on forced-choice items. The semantic differential is essentially like the Likert method but takes up more room on a page and takes a lot more time to write items. The problem with ranking is that many respondents answer such items incorrectly by responding to them like a checklist or leaving out some of the response set. (Don't ever use the ranking method.)

The following discussion is written for the beginner in writing items. It is written to give hints about making goals into criterion variables.

State the goals so that it appears that two or more people could agree when the target person had reached the goal. This can later be tested as reliability but at present the interest is in writing items so don't be too much of a stickler. For example, the goal may be for a psychiatric patient to stop talking to himself. Suppose that the client finds out that this is the goal and talks to himself only when he is not being observed. It should be noted that the behavior has not stopped, but it will not be measured accurately. On the other hand, if the goal is for the person not to talk to himself when someone else is present, the measure is accurate. Such issues should be clarified.

In another example, the goal may for a client to be "free from anxiety." The problem is to make such a goal into a statement in which it could be determined by two or more people that the person was "free of anxiety." The best way is to make as many statements as possible about anxiety and then test the reliability of such statements. Some such statements might be: (1) Do his/her hands shake? (2) Will s/he go out on his/her own? (3) Is s\he fearful? (4) Is s\he afraid of people? etc.

The main point to be made here is that the statement must describe a situation in which some quantitative judgement can be made as to whether the client has reached the goal or not.

The next step is to make as many statements as are feasible to include all the goals which you have stated. The following are some informal rules for writing items for measurement. They are not absolute rules, but the more of them that are broken, the less likely is the reliability of the item. Apply these rules to the statements which you make about the goals.

1. Avoid statements that may be interpreted in more than one way.

2. Avoid statements that are irrelevant to the psychological object under consideration.

3. Avoid statements that refer to the past rather than the present (unless it is specifically the past you are concerned about).

4. Avoid statements that are likely to be endorsed by almost everyone or almost nobody.

5. Select statements that are believed to cover the range of the goal or goals you have outlined.

6. Keep the language of the statements simple, clear, and direct.

7. Statements should be short, rarely exceeding 12 words.

8. Each statement should contain only one complete thought.

9. Statements containing universals such as all, always, none or never, often introduce ambiguity and should be avoided.

10. Such words as only, just, merely, and others of a similar nature should be used with care and moderation in writing statements.

11. Whenever possible statements should be in the form of simple sentences rather than in the form of compound and complex sentences.
12. Avoid the use of words that may not be understood by those who will be using the scale.

13. Avoid the use of double negatives.

Scoring a Likert test:

This scale has been designed so that you can rate a patient on his abilities in certain mental health areas. Please respond to every item. In each case, draw a circle around the letter which represents what you think his abilities are as follows:

SA if you strongly agree with the statement
A if you agree but not as strongly
N if you are neutral
D if you disagree but not too strongly
SD if you strongly disagree

1. The patient hears voices. SA A N D SD

2. The patient shaves himself. SA A N D SD

3. The patient interacts appropriately. SA A N D SD

Face reliability scoring.

Decide which items are positive and which are negative in relation to the goals established. That is, if the goal is for the person to be competent, then the item, "The patient shaves himself" would be positive. On the other hand, "The patient hears voices" is negative. If you have an item which you cannot decide whether it is positive or negative, then you should discard it or change the wording to make it either positive or negative. If you wish to keep the item as it is, you can item analyze the items, and it will be determined. After you have decided whether the item is positive or negative, give values to the letters as follows:

If the item is positive:

SA = 5
A = 4
N = 3
D = 2
SD = 1

If the item is negative:

SA = 1
A = 2
N = 3
D = 4
SD = 5

How many weights should there be on the Likert scale. The true-false item has two

weights. The five point scale above has 5 weights (usually 1, 2, 3, 4, and 5). There are some studies to indicate that reliability improves as the number of weights increase up to about 15. The improvement begins to wane at about 7 or 8. When people are making subjective judgments they tend to give fractional weights when the judgment is between two numbers. For example, when the judgment is either 1 or 2 and the person making the judgment is in between will indicate 1 and a half. This will happen more frequently when there is no middle weight (some people are truly undecided on in the middle--to force them one way or another causes unreliability). It happens in another way: when using a 10 point scale judges will sometimes report 7 and a half. So that this is half way between the "half-way" point and the highest point of the scale. There is some evidence that people can make this "half-way" judgment three time. A nine point scale (0 1 2 3 4 5 6 7 8) allows such a possibility. Four is half-way between 0 and 8, 2 is half-way between 0 and 4, and finally 3 is half-way between 2 and 4. At any rate the 9 point scale (0 through 8) is recommended.

The weights can have different qualitative descriptors. For example, the above Likert scales are based on the strength of agreement (Strongly Agree to Strongly Disagree). The descriptors can be used for various items.

The following descriptors indicate the about of time spent performing an activity.

INSTRUCTIONS: For each item draw a circle around the number that you think best describes the setting according to the following scale.

none of the time		a little of the time		some of the time		a lot of the time		all of the time
0	1	2	3	4	5	6	7	8

When people are in this setting they are:

1. 0 1 2 3 4 5 6 7 8 tense
2. 0 1 2 3 4 5 6 7 8 satisfied

A more detailed descripter of the amount of time:

never	hardly ever	once in a while	little of the time	some of the time	a lot of the time	fre-quent-ly	most of the time	all of the time
0	1	2	3	4	5	6	7	8

Degree of satisfaction

Completely Satisfied		Somewhat Satisfied		Neutral		Somewhat Dissatisfied		Completely Dissatisfied
0	1	2	3	4	5	6	7	8

Degree of Importance.

Not Important		Not Very Important		Somewhat Important		Important		Very Important
0	1	2	3	4	5	6	7	8

These scales have the following format:

INSTRUCTIONS: This scale is designed so that you can indicate how often you experience various emotions. For each item circle the number that best represents how frequently you experience the emotion according to the following scale:

never	rarely	infrequently	occasionally	sometimes	commonly	frequently	usually	always
0	1	2	3	4	5	6	7	8

1. Happiness 0 1 2 3 4 5 6 7 8
2. Sadness 0 1 2 3 4 5 6 7 8
3. Anger 0 1 2 3 4 5 6 7 8
4. Worthwhile 0 1 2 3 4 5 6 7 8

Then each item would have the score of the number circled. For a total score, these numbers would then be added together. This score would be an estimate of the positive emotions of the patients that had been rated.

There are a number of problems with this score. First, two of the items represent positive emotions while two items represents a positive emotions. If you wanted to add the items together to a total test score this issue would need to be resolved. Lets assume that the purpose of the test is to measure positive affect. The items "happy" and "worthwhile" could be added together but the items "sad" and "anger" are not positive emotions. Consequently, the weight on these items

would contribute to negative emotions. A zero (0) on the negative emotion should be changed to an 8 on positive emotion, while a score of eight (8) on negative emotion should be scored zero (0) on positive emotion. This is sometimes referred to as "reversing the item." A two (2) becomes a 6, and a 6 becomes a 2.

The next problem with these weights is that the actual weights are unknown. For example, is a score of 5 on Item #2 worth the same as a score of 5 on Item #3? The above scoring method assumes that it does.

Third, it assumes that different people will agree when they rate the same patient on the same item. For example, reliability assumes that two or more people will rate the same patient the same on item #3.

A fourth problem is that it is assumed that a certain score of 3 would have some kind of meaning. The only way to know this is to compare a certain score with other scores.

These can be solved or at least the error made can be estimated by testing reliability, standardized the best, and weighing the items. However, that is time consuming and you may want to evaluate the program and risk unreliability. And furthermore, you may be able to show validity by accounting for variance later in the program. On the other hand, you may want to check reliability particularly if you have been through the program and your methodology did not account for much of the variance. To check reliability and further standardize the test. If this is not your first time through the system and you accounted for much of the criterion variable, but suspect its validity, then check validity.