

Subha Kalyan
2017.

CHAPTER

9

MEASUREMENT: SCALING, RELIABILITY, VALIDITY

TOPICS DISCUSSED

SCALING TECHNIQUES FREQUENTLY USED

- Rating Scales
 - *Dichotomous Scale*
 - *Category Scale*
 - *Likert Scale*
 - *Semantic Differential Scale*
 - *Numerical Scales*
 - *Itemized Rating Scale*
 - *Fixed or Constant Sum Rating Scale*
 - *Stapel Scale*
 - *Graphic Rating Scale*
 - *Consensus Scale*
- Ranking Scales
 - *Paired Comparisons*
 - *Forced Choice*
 - *Comparative Scale*

GOODNESS OF MEASURES

- Stability
 - *Test-Retest Reliability*
 - *Parallel-Form Reliability*
- Internal Consistency
 - *Split-Half Reliability*
 - *Interitem Consistency Reliability*
- Validity
 - Content Validity
 - *Face Validity*
 - Criterion-Related Validity
 - *Concurrent Validity*
 - *Predictive Validity*
 - Construct Validity

CHAPTER OBJECTIVES

After completing Chapter 9, you should be able to:

1. Know how and when to use the different forms of rating scales and ranking scales.
2. Explain stability and consistency and how they are established.
3. Be conversant with the different forms of validity.
4. Discuss what "goodness" of measures means, and why it is necessary to establish it in research.

Now that we know the four different types of scales that can be used to measure the operationally defined dimensions and elements of a variable, it is necessary to examine the methods of scaling (that is, assigning numbers or symbols) to elicit the attitudinal responses of subjects toward objects, events, or persons. There are two main categories of attitudinal scales (not to be confused with the four different *types of scales*)—the *rating scale* and the *ranking scale*. Rating scales have several response categories and are used to elicit responses with regard to the object, event, or person studied. Ranking scales, on the other hand, make comparisons between or among objects, events, or persons and elicit the preferred choices and ranking among them. Both scales are discussed below.

RATING SCALES

The following rating scales are often used in organizational research:

- Dichotomous scale
- Category scale
- Likert scale
- Numerical scales
- Semantic differential scale
- Itemized rating scale
- Fixed or constant sum rating scale
- Stapel scale
- Graphic rating scale
- Consensus scale

Other scales such as the Thurstone Equal Appearing Interval Scale, and the Multidimensional Scale are less frequently used. We will briefly describe each of the above attitudinal scales.

Dichotomous Scale

The dichotomous scale is used to elicit a Yes or No answer, as in the example below. Note that a nominal scale is used to elicit the response.

Example 9.1 Do you own a car? Yes No

Category Scale

The category scale uses multiple items to elicit a single response as per the following example. This also uses the nominal scale.

Example 9.2 Where in northern California do you reside? North Bay
 South Bay
 East Bay
 Peninsula
 Other

Likert Scale

The Likert scale is designed to examine how strongly subjects agree or disagree with statements on a 5-point scale with the following anchors:

Strongly Disagree	Disagree	Neither Agree Nor Disagree	Agree	Strongly Agree
1	2	3	4	5

The responses over a number of items tapping a particular concept or variable (as per the following example) are then summated for every respondent. This is an interval scale and the differences in the responses between any two points on the scale remain the same.

Example 9.3

Using the preceding Likert scale, state the extent to which you agree with each of the following statements:

My work is very interesting	1	2	3	4	5
I am not engrossed in my work all day	1	2	3	4	5
Life without my work will be dull	1	2	3	4	5

Semantic Differential Scale

Several bipolar attributes are identified at the extremes of the scale, and respondents are asked to indicate their attitudes, on what may be called a semantic

space, toward a particular individual, object, or event on each of the attributes. The bipolar adjectives used, for instance, would employ such terms as Good-Bad; Strong-Weak; Hot-Cold. The semantic differential scale is used to assess respondents' attitudes toward a particular brand, advertisement, object, or individual. The responses can be plotted to obtain a good idea of their perceptions. This is treated as an interval scale. An example of the semantic differential scale follows.

Example 9.4

Responsive	—	—	†	—	—	—	—	Unresponsive
Beautiful	—	—	—	—	—	†	—	Ugly
Courageous	†	—	—	—	—	—	—	Timid

Numerical Scale

The numerical scale is similar to the semantic differential scale, with the difference that numbers on a 5-point or 7-point scale are provided, with bipolar adjectives at both ends, as illustrated below. This is also an interval scale.

Example 9.5

How pleased are you with your new real estate agent?

Extremely										Extremely
Pleased	7	6	5	4	3	2	1			Displeased

Itemized Rating Scale

A 5-point or 7-point scale with anchors, as needed, is provided for each item and the respondent states the appropriate number on the side of each item, or circles the relevant number against each item, as per the examples that follow. The responses to the items are then summated. This uses an interval scale.

Example 9.6 (i)

Respond to each item using the scale below, and indicate your response number on the line by each item.

	1	2	3	4	5
	Very Unlikely	Unlikely	Neither Unlikely Nor Likely	Likely	Very Likely
1. I will be changing my job within the next 12 months.					—
2. I will take on new assignments in the near future.					—
3. It is possible that I will be out of this organization within the next 12 months.					—

Note that the above is a *balanced rating scale* with a *neutral* point.

Example 9.6 (ii) Circle the number that is closest to how you feel for the item below.

Not at All Interested 1	Somewhat Interested 2	Moderately Interested 3	Very Much Interested 4		
How would you rate your interest in changing current organizational policies?		①	②	3	④

This is an *unbalanced rating scale* which does *not* have a neutral point.

The itemized rating scale provides the flexibility to use as many points in the scale as considered necessary (4, 5, 7, 9, or whatever), and it is also possible to use different anchors (e.g., Very Unimportant to Very Important; Extremely Low to Extremely High). When a neutral point is provided, it is a balanced rating scale, and when it is not, it is an unbalanced rating scale.

Research indicates that a 5-point scale is just as good as any, and that an increase from 5 to 7 or 9 points on a rating scale does not improve the reliability of the ratings (Elmore & Beggs, 1975).

The itemized rating scale is frequently used in business research, since it adapts itself to the number of points desired to be used, as well as the nomenclature of the anchors, as is considered necessary to accommodate the needs of the researcher for tapping the variable.

Fixed or Constant Sum Scale

The respondents are here asked to distribute a given number of points across various items as per the example below. This is more in the nature of an ordinal scale.

Example 9.7 In choosing a toilet soap, indicate the importance you attach to each of the following five aspects by allotting points for each to total 100 in all.

Fragrance	—
Color	—
Shape	—
Size	—
Texture of lather	—
Total points	100

Stapel Scale

This scale simultaneously measures both the direction and intensity of the attitude toward the items under study. The characteristic of interest to the study is placed at the center and a numerical scale ranging, say, from + 3 to - 3, on either

side of the item as illustrated below. This gives an idea of how close or distant the individual response to the stimulus is, as shown in the example below. Since this does not have an absolute zero point, this is an interval scale.

Example 9.8

State how you would rate your supervisor's abilities with respect to each of the characteristics mentioned below, by circling the appropriate number.

+3	+3	+3
+2	+2	+2
+1	+1	+1
Adopting Modern Technology	Product Innovation	Interpersonal Skills
-1	-1	-1
-2	-2	-2
-3	-3	-3

Graphic Rating Scale

A graphical representation helps the respondents to indicate on this scale their answers to a particular question by placing a mark at the appropriate point on the line, as in the following example. This is an ordinal scale, though the following example might appear to make it look like an interval scale.

Example 9.9

*On a scale of 1 to 10,
how would you rate
your supervisor?*

10	Excellent
5	All right
1	Very bad

This scale is easy to respond to. The brief descriptions on the scale points are meant to serve as a guide in locating the rating rather than represent discrete categories. The **faces scale**, which depicts faces ranging from smiling to sad (illustrated in Chapter 10), is also a graphic rating scale. used to obtain responses regarding people's feelings with respect to some aspect—say, how they feel about their jobs.

Consensus Scale

Scales are also developed by consensus, where a panel of judges selects certain items, which in its view measure the relevant concept. The items are chosen particularly based on their pertinence or relevance to the concept. Such a consensus scale is developed after the selected items are examined and tested for their

validity and reliability. One such consensus scale is the **Thurstone Equal Appearing Interval Scale**, where a concept is measured by a complex process followed by a panel of judges. Using a pile of cards containing several descriptions of the concept, a panel of judges offers inputs to indicate how close or not the statements are to the concept under study. The scale is then developed based on the consensus reached. However, this scale is rarely used for measuring organizational concepts because of the time necessary to develop it.

Other Scales

There are also some advanced scaling methods such as **multidimensional scaling**, where objects, people, or both, are visually scaled, and a conjoint analysis is performed. This provides a visual image of the relationships in space among the dimensions of a construct.


It is to be noted that usually the Likert or some form of numerical scale is usually the one most frequently used to measure attitudes and behaviors in organizational research.

RANKING SCALES

As already mentioned, **ranking scales** are used to tap preferences between two or among more objects or items (ordinal in nature). However, such ranking may not give definitive clues to some of the answers sought. For instance, let us say there are four product lines and the manager seeks information that would help decide which product line should get the most attention. Let us also assume that 35% of the respondents choose the first product, 25% the second, and 20% choose each of products three and four as of importance to them. The manager cannot then conclude that the first product is the most preferred since 65% of the respondents did not choose that product! Alternative methods used are the *paired comparisons*, *forced choice*, and the *comparative* scale, which are discussed below.

Paired Comparison

The **paired comparison** scale is used when, among a small number of objects, respondents are asked to choose between two objects at a time. This helps to assess preferences. If, for instance, in the previous example, during the paired comparisons, respondents consistently show a preference for product one over products two, three, and four, the manager reliably understands which product line demands his utmost attention. However, as the number of objects to be compared increases, so does the number of paired comparisons. The paired choices for n objects will be $[(n)(n-1)/2]$. The greater the number of objects or stimuli, the greater the number of paired comparisons presented to the respondents, and the greater the respondent fatigue. Hence paired comparison is a good method if the number of stimuli presented is small.

$[(n)(n-1)/2]$


Forced Choice

The **forced choice** enables respondents to rank objects relative to one another, among the alternatives provided. This is easier for the respondents, particularly if the number of choices to be ranked is limited in number.

Example 9.10

Rank the following magazines that you would like to subscribe to in the order of preference, assigning 1 for the most preferred choice and 5 for the least preferred.

Fortune	—
Playboy	—
Time	—
People	—
Prevention	—

Comparative Scale

The **comparative scale** provides a benchmark or a point of reference to assess attitudes toward the current object, event, or situation under study. An example of the use of comparative scale follows.

Example 9.11

In a volatile financial environment, compared to stocks, how wise or useful is it to invest in Treasury bonds? Please circle the appropriate response.

More Useful		About the Same		Less Useful
1	2	3	4	5

In sum, nominal data lend themselves to dichotomous or category scale; ordinal data to any one of the ranking scales—paired comparison, forced choice, or comparative scales; and interval or interval-like data to the other rating scales, as seen from the various examples above. The semantic differential and the numerical scales are, strictly speaking, not interval scales, though they are often treated as such in data analysis.

Rating scales are used to measure most behavioral concepts. Ranking scales are used to make comparisons or rank the variables that have been tapped on a nominal scale.

✓ GOODNESS OF MEASURES

Now that we have seen how to operationally define variables and apply different scaling techniques, it is important to make sure that the instrument that we develop to measure a particular concept is indeed *accurately* measuring the variable, and that in fact, we are *actually* measuring the concept that we set out to measure. This ensures that in operationally defining perceptual and attitudinal

variables, we have not overlooked some important dimensions and elements or included some irrelevant ones. The scales developed could often be imperfect, and errors are prone to occur in the measurement of attitudinal variables. The use of better instruments will ensure more accuracy in results, which in turn, will enhance the scientific quality of the research. Hence, in some way, we need to assess the "goodness" of the measures developed. That is, we need to be reasonably sure that the instruments we use in our research do indeed measure the variables they are supposed to, and that they measure them accurately.

Let us now examine how we can ensure that the measures developed are reasonably good. First an item analysis of the responses to the questions tapping the variable is done, and then the reliability and validity of the measures are established, as described below.

Item Analysis

Item analysis is done to see if the items in the instrument belong there or not. Each item is examined for its ability to discriminate between those subjects whose total scores are high, and those with low scores. In item analysis, the means between the high-score group and the low-score group are tested to detect significant differences through the *t*-values (see Module at the end of the book for explanation of *t*-tests). The items with a high *t*-value (test which is able to identify the highly discriminating items in the instrument) are then included in the instrument. Thereafter, tests for the reliability of the instrument are done and the validity of the measure is established.

Very briefly, reliability tests *how consistently* a measuring instrument measures whatever concept it is measuring. Validity tests how well an instrument that is developed measures the *particular concept* it is intended to measure. In other words, validity is concerned with whether we measure the right concept, and reliability with stability and consistency of measurement. Validity and reliability of the measure attest to the scientific rigor that has gone into the research study. These two criteria will now be discussed. The various forms of reliability and validity are depicted in Figure 9.1.

RELIABILITY

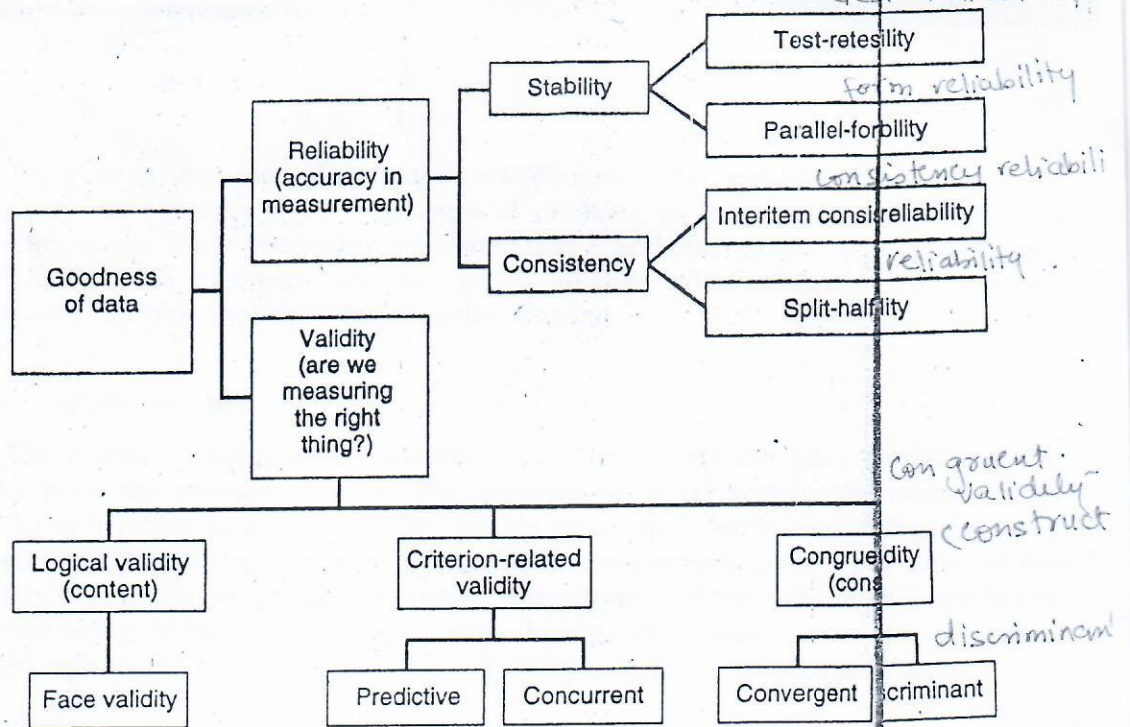
The reliability of a measure indicates the extent to which it is without bias (error free) and hence ensures consistent measurement across time and across the various items in the instrument. In other words, the reliability of a measure is an indication of the stability and consistency with which the instrument measures the concept and helps to assess the "goodness" of a measure.

Stability of Measures

The ability of a measure to remain the same over time—despite uncontrollable testing conditions or the state of the respondents themselves—is indicative of its

Figure 9.1

Testing Goodness of Measures: Forms of Reliability and Validity.



stability and low vulnerability to changes in the situation. This attests to its "goodness" because the concept is stably measured, no matter when it is done. Two tests of stability are test-retest reliability and parallel-form reliability.

Test-Retest Reliability

The reliability coefficient obtained with a repetition of the same measure on a second occasion is called test-retest reliability. That is, when a questionnaire containing some items that are supposed to measure a concept is administered to a set of respondents now, and again to the same respondents several weeks to 6 months later, then the correlation between the scores obtained at the two different times from one and the same set of respondents is called the test-retest coefficient. The higher it is, the better the test-retest reliability, and consequently, the stability of the measure across time.

Parallel-Form Reliability

When responses on two comparable sets of measures tapping the same construct are highly correlated, we have parallel-form reliability. Both forms have similar items and the same response format, the only changes being the wordings and

the order or sequence of the questions. What we try to establish here is the error variability resulting from wording and ordering of the questions. If two such comparable forms are highly correlated (say 8 and above), we may be fairly certain that the measures are reasonably reliable, with minimal error variance caused by wording, ordering, or other factors.

Internal Consistency of Measures ✓

The internal consistency of measures is indicative of the homogeneity of the items in the measure that tap the construct. In other words, the items should "hang together as a set," and be capable of independently measuring the same concept so that the respondents attach the same overall meaning to each of the items. This can be seen by examining if the items and the subsets of items in the measuring instrument are correlated highly. Consistency can be examined through the inter-item consistency reliability and split-half reliability tests.

Interitem Consistency Reliability ✓

This is a test of the consistency of respondents' answers to all the items in a measure. To the degree that items are independent measures of the same concept, they will be correlated with one another. The most popular test of interitem consistency reliability is the Cronbach's coefficient alpha (Cronbach's alpha; Cronbach, 1946), which is used for multipoint-scaled items, and the Kuder-Richardson formulas (Kuder & Richardson, 1937), used for dichotomous items. The higher the coefficients, the better the measuring instrument.

Split-Half Reliability ✓

Split-half reliability reflects the correlations between two halves of an instrument. The estimates would vary depending on how the items in the measure are split into two halves. Split-half reliabilities could be higher than Cronbach's alpha only in the circumstance of there being more than one underlying response dimension tapped by the measure and when certain other conditions are met as well (for complete details, refer to Campbell, 1976). Hence, in almost all cases, Cronbach's alpha can be considered a perfectly adequate index of the interitem consistency reliability.

It should be noted that the consistency of the judgment of several raters on how they view a phenomenon or interpret some responses is termed *interrater reliability*, and should not be confused with the reliability of a measuring instrument. As we had noted earlier, interrater reliability is especially relevant when the data are obtained through observations, projective tests, or unstructured interviews, all of which are liable to be subjectively interpreted.

It is important to note that reliability is a necessary but not sufficient condition of the test of goodness of a measure. For example, one could very reliably measure a concept establishing high stability and consistency, but it may not be the concept that one had set out to measure. **Validity** ensures the ability of a scale to measure the intended concept. We will now discuss the concept of validity.

VALIDITY ✓

We examined earlier, in Chapter 7, the terms *internal validity* and *external validity* in the context of experimental designs. That is, we were concerned about the issue of the authenticity of the cause-and-effect relationships (internal validity), and their generalizability to the external environment (external validity). We are now going to examine the validity of the measuring instrument itself. That is, when we ask a set of questions (i.e., develop a measuring instrument) with the hope that we are tapping the concept, how can we be reasonably certain that we are indeed measuring the concept we set out to do and not something else? This can be determined by applying certain validity tests.

Several types of validity tests are used to test the goodness of measures and writers use different terms to denote them. For the sake of clarity, we may group validity tests under three broad headings: **content validity**, **criterion-related validity**, and **construct validity**.

Content Validity ✓

Content validity ensures that the measure includes an adequate and representative set of items that tap the concept. The more the scale items represent the domain or universe of the concept being measured, the greater the content validity. To put it differently, content validity is a function of how well the dimensions and elements of a concept have been delineated.

A panel of judges can attest to the content validity of the instrument. Kidder and Judd (1986) cite the example where a test designed to measure degrees of speech impairment can be considered as having validity if it is so evaluated by a group of expert judges (i.e., professional speech therapists).

Face validity is considered by some as a basic and a very minimum index of content validity. Face validity indicates that the items that are intended to measure a concept, do on the face of it look like they measure the concept. Some researchers do not see it fit to treat face validity as a valid component of content validity.

Criterion-Related Validity ✓

can be established by

Criterion-related validity is established when the measure differentiates individuals on a criterion it is expected to predict. This can be done by establishing *concurrent validity* or *predictive validity*, as explained below.

Concurrent validity is established when the scale discriminates individuals who are known to be different; that is, they should score differently on the instrument as in the example that follows.

Example 9.12

If a measure of work ethic is developed and administered to a group of welfare recipients, the scale should differentiate those who are enthusiastic about accepting a job and glad of an opportunity to be off welfare, from those who would

not want to work even when offered a job. Obviously, those with high work ethic values would not want to be on welfare and would yearn for employment to be on their own. Those who are low on work ethic values, on the other hand, might exploit the opportunity to survive on welfare for as long as possible, deeming work to be a drudgery. If both types of individuals have the same score on the work ethic scale, then the test would *not* be a measure of work ethic, but of something else.

Predictive validity indicates the ability of the measuring instrument to differentiate among individuals with reference to a future criterion. For example, if an aptitude or ability test administered to employees at the time of recruitment is to differentiate individuals on the basis of their future job performance, then those who score low on the test should be poor performers and those with high scores good performers.

Construct Validity

Construct validity testifies to how well the results obtained from the use of the measure fit the theories around which the test is designed. This is assessed through **convergent** and **discriminant** validity, which are explained below.

Convergent validity is established when the scores obtained with two different instruments measuring the same concept are highly correlated.

Discriminant validity is established when, based on theory, two variables are predicted to be uncorrelated, and the scores obtained by measuring them are indeed empirically found to be so.

Validity can thus be established in different ways. Published measures for various concepts usually report the kinds of validity that have been established for the instrument, so that the user or reader can judge the "goodness" of the measure. Table 9.1 summarizes the kinds of validity discussed here.

Some of the ways in which the above forms of validity can be established are through (1) *correlational analysis* (as in the case of establishing concurrent and predictive validity or convergent and discriminant validity), (2) factor analysis, a multivariate technique that would confirm the dimensions of the concept that have been operationally defined, as well as indicate which of the items are most appropriate for each dimension (establishing construct validity), and (3) the multitrait, multimethod matrix of correlations derived from measuring concepts by different forms and different methods, additionally establishing the robustness of the measure.

In sum, the **goodness of measures** is established through the different kinds of validity and reliability depicted in Figure 9.1. The results of any research can only be as good as the measures that tap the concepts in the theoretical framework. We need to use well-validated and reliable measures to ensure that our research is scientific. Fortunately, measures have been developed for many important concepts in organizational research and their psychometric properties (i.e., the reliability and validity) established by the developers. Thus, researchers can use the instruments already reputed to be "good," rather than laboriously develop their own measures. When using these measures, however, researchers

Table 9.1
Types of Validity

Validity	Description
Content validity	Does the measure adequately measure the concept?
Face validity	Do "experts" validate that the instrument measures what its name suggests it measures?
Criterion-related validity	Does the measure differentiate in a manner that helps to predict a criterion variable?
Concurrent validity	Does the measure differentiate in a manner that helps to predict a criterion variable currently?
Predictive validity	Does the measure differentiate individuals in a manner as to help predict a future criterion?
Construct validity	Does the instrument tap the concept as theorized?
Convergent validity	Do two instruments measuring the concept correlate highly?
Discriminant validity	Does the measure have a low correlation with a variable that is supposed to be unrelated to this variable?

should cite the source (i.e., the author and reference) so that the reader can seek more information if necessary.

It is not unusual that two or more equally good measures are developed for the same concept. For example, there are several different instruments for measuring the concept of job satisfaction. One of the most frequently used scales for the purpose, however, is the Job Descriptive Index (JDI) developed by Smith, Kendall, and Hulin (1969). When more than one scale exists for any variable, it is preferable to use the measure that has better reliability and validity and is also more frequently used.

At times, we may also have to adapt an established measure to suit the setting. For example, a scale that is used to measure job performance, job characteristics, or job satisfaction in the manufacturing industry may have to be modified slightly to suit a utility company or a health care organization. The work environment in each case is different and the wordings in the instrument may have to be suitably adapted. However, in doing this, we are tampering with an established scale, and it would be advisable to test it for the adequacy of the validity and reliability afresh.

A sample of a few measures used to tap some frequently researched concepts in the management and marketing areas is provided in the Appendix to this chapter.

SUMMARY

In this chapter, we saw what kinds of attitude rating scales and ranking scales can be used in developing instruments after a concept has been operationally defined. We also discussed how the goodness of measures is established by means of item analysis, and reliability and validity tests. We also noted that the Likert scale and other types of interval-type