

# **An In-Depth Examination of Validity Assessment: Exploring Diverse Methodologies and Dimensions of Validity in Social Research Studies**

## **Abstract**

Validity refers to the extent to which a test accurately measures what it claims to measure. Validity is a matter of degree. The study is done to know the various methods of validity. Validity is a fundamental aspect of research quality, which is why it is important to assess the various methods of validity. Descriptive research was used for the study. It was examined that face validity refers to an indicator that seems to measure its construct “on its face”. Content validity is a form of judgmental validity, that involves determining if a test adequately represents the content domain. Criterion-related validity methods include concurrent and predictive approaches. Construct validity assesses how well theoretical ideas are translated into measurable variables. Face validity involves examining whether a measurement appears suitable based on its operationalization on the face of it. Lawshe's content validity ratio quantifies expert agreement on the essentiality of individual items in a test or scale. In contrast, the content validity index considers overall expert agreement on the representativeness of items collectively within a test or scale. It considers the agreement among experts regarding the content validity of the entire set of items. Cohen's Kappa is used to assess the level of agreement between two or more raters or observers when coding or categorizing data. The kappa coefficient is particularly useful when chance agreement among raters is possible. Item Objective Congruence evaluates the efficacy of items in measuring one or more objectives, accomplished by an unbiased expert panel provides evidence of content validity. The most significant assessment at this stage is determining if items and objectives are congruent. Concurrent and predictive validity are assessed using Pearson's correlation and Spearman rank correlation coefficients respectively. The Multi-trait Multimethod Matrix (MTMM), evaluates the construct validity of measures. Validity is essential for ensuring credible and meaningful research findings by enhancing the trustworthiness of the measuring instruments.

**Keywords:** Validity, Lawshe, CVR, CVI, Kappa, IOC, Pearson, Spearman rank, MTMM

## **Introduction**

“Validity refers to the extent to which a test accurately measures what it claims to measure” (Kelley, 1927). “The commonest definition of validity is epitomized by the question: Are we measuring what we think we are measuring? The emphasis in this question is on what is being measured. For example, a teacher has constructed a test to measure understanding of scientific

procedures and has included only factual items about scientific procedures in the test. The test is not valid, because while it may be reliable in measuring the students' factual knowledge of scientific procedures, it does not measure their understanding of such procedures. In other words, it may measure what it measures quite well, but it does not measure what the teacher intended it to measure". (Kerlinger, 1986). "One validates, not a test, but an interpretation of data arising from a specified procedure" (Cronbach, 1971). "One does not assess the validity of an indicator but rather the use to which it is being put. For example, an intelligence test may be valid for assessing the native intellectual potential of students, but it would not necessarily be valid for other purposes, such as forecasting their level of income during adulthood" (Nunnally, 1978). "Just as reliability is a matter of degree, also is validity. Thus, the objective of attaining a perfectly valid indicator-one that represents the intended, and only the intended, concept is unachievable. Instead, validity is a matter of degree, not an all-or-none property. Moreover, just because an indicator is quite reliable, this does not mean that it is also relatively valid" (Nunnally, 1978). Reliability is an empirical issue, focusing on the performance of empirical measures. Validity, in contrast, is usually more of a theoretically oriented issue because it inevitably raises the question, "valid for what purpose?" (Carmines & Zeller, 2007). "Quantitative research has the advantages of reliability and generalizability of the findings, and advances in data collection and analysis methods have increased reliability and generalizability whereas qualitative research has the advantage of validity, which refers to how accurately or appropriately a phenomenon was measured" (Pyo *et al.*, 2023).

## **Review of Literature**

Herche (1992) found that "the predictive validity of Shimp and Sharma's (1987) CETSCALE is tested in a nationwide mail survey. The scale is shown to be a much stronger predictor of import buying behavior".

Yaghmaie (2003) concluded that "for content validity two judgments are necessary: the measurable extent of each item for defining the traits and the set of items that represents all aspects of the traits. It was found that out of 38 items, those with CVI over 0.75 remained and the rest were discarded resulting in a 25-item scale".

Dellinger (2005) placed "construct validity, or the meaning of measures, as the central focus or core of validity encompassing all forms of validity evidence. Any information provided as valid evidence through traditional means serves to bolster or negate arguments that data from a tool, process, observation, instrument, etc., are valid indicators of the construct of interest".

Watson *et al.* (2008) reported the validity of the Inventory of Depression and Anxiety Symptoms (IDAS; Watson *et al.*, 2007) in two samples (306 college students, and 605 psychiatric patients). The IDAS scales showed strong convergent validity about parallel interview-based scores on the Clinician Rating version of the IDAS (IDAS-CR); the mean convergent correlations were .51

and .62 in the student and patient samples, respectively.

Handage and Chander (2021) found “the content validity is assessed in two rounds, in the first round the Content Validity Ratio (CVR) is computed and the items with a CVR critical value of 0.407 or above were accepted and subjected to the computation of Content Validity Index (CVI) in the second round. Modified kappa values were determined to overcome the chancefactor. The Item-Content Validity Index (I-CVI) values were computed to finalize the items in the final scale while the Scale-Content Validity Index (S-CVI) was determined to check for the stability of each dimension as well as the scale as a whole. The CVI value above 0.78 is considered fit for the scale”.

Priyadarshani *et al.* (2021) concluded that the Item discrimination index is used for the discriminant validity, all items could discriminate between the high and low groups significantly ( $p < 0.01$ ), and the final test compromised of a total of 12 items with discriminant validity.

Singh *et al.* (2023) reported that “the concurrent validity of the Social Cohesiveness Rating Scale (SCRS) is examined by administering Spearman rank-order correlation between the mean score of the two raters completing each scale with the mean score of the other scale. The correlation between SCRS and Group Cohesion Scale (GCS) is 0.75 ( $p < 0.001$ ), and this high significant correlation implies a good consideration of the validity of the SCRS”.

## **Methodology**

The research is descriptive. The study is conceptual and based on secondary data collected from journals, books, websites, and previous studies. The paper focuses on the different methods of assessing the validity of the research instruments/tools like scale, test, etc.

### **2. Types of validity (Test, Scale, and indexes)**

- i. Face validity**
- ii. Content validity**
- iii. Criterion Related validity**
- iv. Construct validity**

#### **2.1 Face validity**

In face validity, you look at the operationalization and see whether "on its face" seems like a good translation of the construct. This is probably the weakest way to try to demonstrate construct validity. “We can improve the quality of face validity assessment considerably by making it more systematic. For instance, if you are trying to assess the face validity of a math ability measure, it would be more convincing if you sent the test to a carefully selected sample of experts on math ability testing and they all reported back with the judgment that your measure appears to be a good measure of math ability” (Jain, 2005).

#### **2.2 Content validity**

Content validity also known as logical validity is the representativeness or the sampling

adequacy of the measuring instrument. It asks the question "To what extent the content of the test/scale is representative of the domain of the psychological object under investigation?" For example, when a trainer is interested in measuring the effect of training on the knowledge level of the farmer, he will initially construct a knowledge test and subject it to content validity to determine to what extent the questions/items in the knowledge test are representative of the training course content. **Content validity is useful in a cognitive task that does not require an understanding of the phenomena under examination (Wang *et al.*, 2023).**

### **2.2.1 Preconditions for Content Validation**

1. Each item must be judged separately for its presumed relevance to the property being measured.
2. Judgement should be done by experts who have experience in studying Job serving/experiencing the phenomenon under investigation.
3. It is not sufficient if all the items are measuring what they were supposed to measure.
4. It should be ensured that the range items are capable of catching every dimension of the construct under investigation

Content validity is judgmental validity and it is not a comprehensive measure of validity. Hence in most cases, content validity alone cannot be used as a measure of validity. It must be combined with other methods of validation for better results (Patil & Aditya, 20016).

### **2.3 Criterion-Related Validity**

“Criterion validity is the idea that a valid test should relate closely to other measures of the same theoretical construct. A valid test of intelligence should correlate highly with other intelligence tests. It should also correlate with behaviors that are considered to require intelligence, such as doing well in school” (White and McBurney, 2012). Criterion-related validity is of two types one is concurrent validity and the second is predictive validity. Concurrent validity is used generally when a test is constructed to replace an existing test that measures the same construct. Predictive validity is generally used to demonstrate that a given measuring instrument makes accurate predictions about the construct it is supposed to measure. In the case of predictive validity, one does not care about what is being measured by instrument but one cares for its predictive ability. It is very concerned about the practical problem and outcomes.

### **2.4 Construct Validity (Convergent and Discriminant Validity)**

The construct validity approach is more complex than other forms of validity. White and McBurney (2012) defined “construct validity as the test should measure whatever theoretical construct it supposedly tests, and not something else. There are several ways to determine whether a test generates data that has construct validity”.

1. The test should measure whatever theoretical construct it supposedly tests, and not

something else. For example, a test of leadership ability should not test extraversion.

2. A test that has construct validity should measure what it intends to measure but not measure theoretically unrelated constructs. For example, a test of musical aptitude should not require too much reading ability.
3. A test should prove useful in predicting results related to the theoretical concepts it is measuring. For example, a test of musical ability should predict who will benefit from taking music lessons, should differentiate groups who have chosen music as a career from those who haven't should relate to other tests of musical ability, and so on.

There are two types of construct validity— 'convergent validity' and 'divergent validity' (or discriminant validity). In convergent validity, the degree to which the operationalization is similar to (converges on) other operationalizations that it theoretically should be similar to. In discriminant validity, the degree to which the operationalization is not similar to (diverges from) other operationalizations that it theoretically should not be similar to (Jain, 2005).

### 3. Methods for Assessing various forms of Validity

#### 3.1 Lawshe's Content Validity Ratio

The Content Validity Ratio (CVR) is used to assess the content validity of individual items in a test or scale. It was introduced by Lawshe in 1975. The CVR determines whether each item is essential for measuring the construct being assessed. To calculate the CVR, a panel of experts rates each item as either "essential" or "not essential" for the test.

The CVR formula is as follows:  $CVR = (ne - (N/2)) / (N/2)$

Where:

ne = number of experts who rate the item as "essential"

N = total number of experts

The CVR ranges from -1 to +1, and a positive value indicates that the item is considered essential by at least some experts. The commonly used criterion is that an item is considered essential if the CVR is greater than or equal to 0.62. The CVR (content validity ratio) is a linear transformation of a proportional level of agreement on how many 'experts' within a panel rate an item 'essential'. "It could readily be seen whether the level of agreement amongst panel members is greater than 50%. CVR values range between -1 (perfect disagreement) and +1 (perfect agreement) with CVR values above zero indicating that over half of the panel members agree on an item essential" (Ayre and Scally, 2014). The final CVR computation is dependent on the number of judges or experts who assess the items. Any item that is perceived by more than half of experts to be essential has some degree of content validity.

**Table. 1: Minimum value of CVR (Lawshe, 1975)**

| No. of judges | Minimum value |
|---------------|---------------|
|---------------|---------------|

|    |     |
|----|-----|
| 5  | .99 |
| 6  | .99 |
| 7  | .99 |
| 8  | .78 |
| 9  | .75 |
| 10 | .62 |

CVR is calculated in the following way: 
$$CVR = \frac{ne - \frac{N}{2}}{\frac{N}{2}}$$

Where;

CVR =content validity ratio

ne = Number of panel members indicating an item ‘essential’

N = Number of panel members

**Example:**

Gumus and Kumul (2023) took “the opinions of eight experts to determine whether the items were clear. The comments of these experts on the scale items were received. In line with the opinions of the experts, the items were re-examined and necessary adjustments were made. Items that were deemed unsuitable were removed from the scale, appropriate ones from the suggested items were added, and items that needed to be adjusted were updated in line with the expert opinions”. As a result, a draft scale consisting of a total of 69 items with a 5-point Likert scale was prepared.

**Table. 2: Calculation of content validity ratio (CVR)**

| Item | Expert 1 | Expert 2 | Expert 3 | Expert 4 | Expert 5 | CVR |
|------|----------|----------|----------|----------|----------|-----|
| 1    | 1        | 1        | 0        | 0        | 1        | 0.2 |
| 2    | 1        | 1        | 1        | 1        | 1        | 1.0 |
| 3    | 1        | 1        | 1        | 0        | 1        | 0.6 |
| 4    | 1        | 1        | 1        | 0        | 1        | 0.6 |
| 5    | 1        | 1        | 1        | 1        | 0        | 0.6 |

CVR of first item:

$$CVR = \frac{3-5/2}{5/2} \quad CVR = 0.2$$

### 3.2 Content Validity Index (CVI)

The CVI is a measure used to assess the relevance and representativeness of items collectively within a test or scale. It considers the agreement among experts regarding the content validity of the entire set of items. The CVI is calculated for all individual items (I-CVI) and the overall scale (S-CVI). For CVI, the panel of experts is asked to rate each scale item in terms of its

relevance to the underlying construct. A 4-point scale is used to avoid a neutral point. The four points used along the item rating continuum were 1= not relevant, 2 = somewhat relevant, 3 = quite relevant, and 4 = highly relevant.

$$\text{I-CVI} = (\text{number of experts giving a rating of 3 or 4}) / (\text{total number of experts})$$

The S-CVI ranges from 0 to 1, with higher values indicating greater content validity. A commonly used threshold for an acceptable level of content validity is an S-CVI of 0.80 or higher, meaning that at least 80% of the experts agree on the relevance or representativeness of each item (Shrotryia and Dhanda, 2019).

**Table. 3: Computation of content validity index (CVI)**

| Item | E1 | E2 | E3 | E4 | E5 | E6 | A | I-CVI |
|------|----|----|----|----|----|----|---|-------|
| 1    | 1  | 1  | 1  | 1  | 1  | 1  | 6 | 1.00  |
| 2    | 1  | 1  | 1  | 0  | 1  | 1  | 5 | 0.83  |
| 3    | 1  | 1  | 1  | 1  | 1  | 1  | 6 | 1.00  |
| 4    | 1  | 0  | 1  | 1  | 0  | 1  | 4 | 0.83  |
| 5    | 1  | 1  | 1  | 1  | 1  | 1  | 6 | 1.00  |

I-CVI: item content validity index, S-CVI: scale content validity index, A: no. in agreement

I-CVI = Rating given by the expert / total number of experts

$$\text{I-CVI} = 6/6 = 1$$

I-CVI should be 1.00 in case of five or fewer judges and case of six or more judges; I-CVI should not be less than 0.78.

$$\text{S-CVI} = \Sigma (\text{I-CVI}) / n$$

$$\text{S-CVI} = 4.49 / 5$$

$$\text{S-CVI (Average)} = 0.898 \text{ (accepted).}$$

where n is the total number of items

It is recommended that a minimum S-CVI should be 0.8 for reflecting content validity.

### 3.2.1 Kappa Statistic coefficient / Cohen's kappa (Jacob Cohen, 1968).

The Kappa coefficient, also known as Cohen's kappa, was developed by Jacob Cohen, an American statistician and psychologist. Cohen introduced the kappa statistic as a measure of inter-rater agreement or reliability for categorical variables, it has been widely used in various fields to assess the level of agreement between two or more raters or observers when coding or categorizing data. The kappa coefficient is particularly useful when chance agreement among raters is possible.

The Kappa coefficient ensures a better understanding of content validity as it removes any random chance agreement. Kappa statistic is a consensus index of interrater agreement that supplements CVI to ensure that the agreement among experts is beyond chance. Computation of

Kappa Statistic requires the calculation of the probability of chance agreement  $P_c$ .

Kappa statistic is then calculated as  $K = (I-CVI - P_c) / (1 - P_c)$ . Evaluation criteria for Kappa are that values above 0.74, between 0.6 and 0.74, and the ones between 0.4 and 0.59 are considered to be excellent, good, and fair, respectively

**Table. 4: Estimation of kappa statistic coefficient**

| Item | E1 | E2 | E3 | E4 | E5 | E6 | A | I-CVI | Pc       | K        |
|------|----|----|----|----|----|----|---|-------|----------|----------|
| 1    | 1  | 1  | 1  | 1  | 1  | 1  | 6 | 1.00  | 0.015625 | 1.00     |
| 2    | 1  | 1  | 1  | 0  | 1  | 1  | 5 | 0.83  | 0.09375  | 0.81     |
| 3    | 1  | 1  | 1  | 1  | 1  | 1  | 6 | 1.00  | 0.015625 | 1.00     |
| 4    | 1  | 0  | 1  | 1  | 0  | 1  | 4 | 0.66  | 0.9375   | Negative |
| 5    | 1  | 1  | 1  | 1  | 1  | 1  | 6 | 1.00  | 0.015625 | 1.00     |

$P_c$  = probability of chance agreement,

$K$  = kappa statistic

$$P_c = [N! / A! (N - A)!] \times 0.5^N$$

Where;

$N$  = number of experts in the panel,

$A$  = number of experts in the panel who agree that the item is relevant.

$$P_c = [(6! / 6!) (6-6)!] * 0.5^6$$

$$P_c = [(6*5*4*3*2*1 / 6*5*4*3*2*1) (0)!] * 0.5^6$$

$$P_c = (1*1) * 0.015625$$

$$P_c = 0.015625$$

Kappa statistic coefficient for the first statement

$$K = (I-CVI - P_c) / (1 - P_c)$$

Where;

I-CVI = Item Content Validity Ratio

$P_c$  = Probability of chance agreement

$$K = (1 - 0.015625) / (1 - 0.015625)$$

$$K = 1$$

### 3.3 Item Objective Congruence Index

An evaluation of the efficacy of items in measuring one or more objectives, accomplished by an unbiased expert panel provides evidence of content validity. The most significant assessment at this stage is determining if items and objectives are congruent. The remaining item analyses are meaningless if there is insufficient proof that the items are measuring what they are supposed to measure. The index of item-objective congruence (IOC) introduced by Rovinelli & Hambleton (1977) is one method to quantitatively measure content experts' judgments of items to evaluate the

fit between test items and the table of specifications. “IOC is a process in which SMEs (Subject Matter Experts) rate individual items on the degree to which they agree or do not agree with the specific objectives listed by the test developer” (Turner *et al.*, 2003). Accordingly, an expert evaluates each item by giving a rating of +1 for clearly measuring the objective, -1 for not measuring, or 0 for the unclear objective. After the experts rate the items, the results are calculated to create the indices of IOC for each item on each objective. The choice of a cutoff score for this index to separate "good" from "bad" items can be based on some absolute standard relating to specific proportions of perfect ratings for the items. For example, if one-half of the content specialists judged an item to be a perfect match to an objective, while the others were not able to make a decision, the computed value of the index would be .50. Thus, test constructors obtaining I' values of .50 would know that at a minimum, at least 50 percent of the content specialists gave a perfect rating to the item.

In a condition where an item measures more than one objective, the multidimensional item formula simplified by Crocker and Aligna is utilized to evaluate the similarity between an item and a set of objectives. The formula for the multidimensional item given by Crocker and Aligna is as follows:

$$I'_{ik} = (M-1) S_{ik} - S'_{ik} / 2N - (M-1)$$

Where;

$I_{ik}$  is the index of item-objective congruence for item  $i$  and objective  $k$

$M$  is the number of objectives

$N$  is the number of judges

$S_{ik}$  is the sum of the item ratings assigned to objective  $k$

$S'_{ik}$  is the sum of the item ratings assigned to all objectives, except objective  $k$

To illustrate the calculation of the index of item-objective congruence, consider the following example for Item 1 of a test with five objectives or domains. Four judges have been asked to rate each item relative to each of the domains. Their ratings appear in the Table below.

**Table. 5: Item objective congruence for item 1 with 5 objectives**

| Judge | Objectives |    |    |    |    |
|-------|------------|----|----|----|----|
|       | 1          | 2  | 3  | 4  | 5  |
| A     | -1         | +1 | -1 | 0  | -1 |
| B     | -1         | +1 | -1 | -1 | -1 |
| C     | -1         | +1 | -1 | -1 | -1 |

|                      |    |    |    |    |    |
|----------------------|----|----|----|----|----|
| <b>D</b>             | -1 | +1 | -1 | +1 | -1 |
| <b>S<sub>i</sub></b> | -4 | +4 | -4 | -1 | -4 |

For this example, the item-objective congruence for Item 1 and Objective 2 is calculated. Therefore,

$$I_{ik} = I_{12}; \quad M = 5; \quad N = 4; \quad S_{12} = 4; \quad \text{and} \quad S'_{12} = -13.$$

Substitution of the above values into the formula for the index yields:

$$I_{12} = \frac{(5-1)(4) - (-13)}{2(4)(5-1)}$$

$$I_{12} = 0.906$$

### 3.4 Pearson's Correlation for Concurrent Validity

The Pearson correlation measures the strength of the linear relationship between two variables. It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and + 1 meaning a total positive correlation.

Pearson's correlation

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

Where:

r = Pearson correlation coefficient

x = Values in the first set of data

y = Values in the second set of data

n = Total number of values.

**Table. 6: Computation of Person's correlation for concurrent validity**

| <b>Respondents</b> | <b>Newley Constructed test (x)</b> | <b>Standardized Test (y)</b> | <b>x<sup>2</sup></b> | <b>y<sup>2</sup></b> | <b>xy</b> |
|--------------------|------------------------------------|------------------------------|----------------------|----------------------|-----------|
| <b>1.</b>          | 8                                  | 7                            | 64                   | 49                   | 56        |
| <b>2.</b>          | 6                                  | 7                            | 36                   | 49                   | 42        |
| <b>3.</b>          | 8                                  | 9                            | 64                   | 81                   | 72        |
| <b>4.</b>          | 7                                  | 7                            | 49                   | 49                   | 49        |
| <b>5.</b>          | 9                                  | 9                            | 81                   | 81                   | 81        |
| <b>6.</b>          | 5                                  | 6                            | 25                   | 36                   | 30        |
| <b>7.</b>          | 8                                  | 8                            | 64                   | 64                   | 64        |
| <b>8.</b>          | 6                                  | 7                            | 36                   | 49                   | 42        |
| <b>9.</b>          | 9                                  | 9                            | 81                   | 81                   | 81        |
| <b>10.</b>         | 5                                  | 5                            | 25                   | 25                   | 25        |

|  |               |               |                  |                  |                 |
|--|---------------|---------------|------------------|------------------|-----------------|
|  | $\sum x = 71$ | $\sum y = 74$ | $\sum x^2 = 525$ | $\sum y^2 = 564$ | $\sum xy = 542$ |
|--|---------------|---------------|------------------|------------------|-----------------|

**Calculating the values**

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

$$r = \frac{10(542) - (71)(74)}{\sqrt{[10 \cdot 525 - (71)^2][10 \cdot 564 - (74)^2]}}$$

$$r = 0.89$$

➤ The newly constructed test has high concurrent validity.

### 3.5 Spearman's Rank Correlation for Predictive Validity

Spearman's rank correlation measures the strength and direction of association between two ranked variables. It gives the measure of monotonicity of the relation between two variables i.e., how well the relationship between two variables could be represented using a monotonic function.

Spearman's rank coefficient:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Where;

$\rho$  = Spearman rank correlation

$d_i$  = difference between the two ranks of each observation

$n$  = number of observations

**Table. 7: Estimation of Spearman rank correlation for predictive validity**

| Applicants | Rank in Aptitude test | Rank in Job performance | Difference in ranks (d) | $d^2$ |
|------------|-----------------------|-------------------------|-------------------------|-------|
| 1.         | 6                     | 5                       | 1                       | 1     |
| 2.         | 3                     | 2                       | 1                       | 1     |
| 3.         | 1                     | 3                       | -2                      | 4     |
| 4.         | 8                     | 6                       | 2                       | 4     |
| 5.         | 4                     | 7                       | -3                      | 9     |
| 6.         | 2                     | 1                       | 1                       | 1     |
| 7.         | 10                    | 9                       | 1                       | 1     |
| 8.         | 9                     | 8                       | 1                       | 1     |

|     |   |    |    |                   |
|-----|---|----|----|-------------------|
| 9.  | 5 | 4  | 1  | 1                 |
| 10. | 7 | 10 | -3 | 9                 |
|     |   |    |    | $\sum d_i^2 = 32$ |

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$$\rho = 1 - \frac{6 * 32}{10(10^2 - 1)}$$

$$\rho = 0.806$$

- Predictor = Aptitude test
- Criterion = Job performance
- The test has high predictive validity

### 3.6 Multi-trait Multi-method Matrix (MTMM)

The Multi-trait Multimethod Matrix (MTMM) is an approach to assessing the construct validity of a set of measures in a study. It was developed in 1959 by Campbell and Fiske. Along with the MTMM, they also introduced two new types of validity, convergent and discriminant as subcategories of construct validity. We can assess both convergent and discriminant validity using the MTMM. To be able to claim that your measures have construct validity, you have to demonstrate both convergence and discrimination.

To construct an MTMM, you need to arrange the correlation matrix by concepts within methods. The Table shows an MTMM for three concepts (traits A, B, and C) each of which is measured with three different methods (1, 2, and 3) Note that you lay the matrix out in blocks by method. Essentially, the MTMM is just a correlation matrix between your measures.

The MTMM idea provided an operational methodology for assessing construct validity. In one matrix, it was possible to examine both convergent and discriminant validity simultaneously. By its inclusion of methods on an equal footing with traits, Campbell and Fiske stressed the importance of looking for the effects of how we measure in addition to what we measure. And, MTMM provided a rigorous framework for assessing construct validity.

**Table. 8: MMTM matrix for three traits and three methods (Schmitt and Stults, 1986;**

O'Leary-Kelly and Vokurka, 1998)

|                     |                | Method <sub>1</sub>               |                                   |                                   | Method <sub>2</sub>               |                                   |                                   | Method <sub>3</sub>               |                                   |                                   |
|---------------------|----------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|
|                     | Traits         | A <sub>1</sub>                    | B <sub>1</sub>                    | C <sub>1</sub>                    | A <sub>2</sub>                    | B <sub>2</sub>                    | C <sub>2</sub>                    | A <sub>3</sub>                    | B <sub>3</sub>                    | C <sub>3</sub>                    |
| Method <sub>1</sub> | A <sub>1</sub> | <b>A<sub>1</sub>A<sub>1</sub></b> |                                   |                                   |                                   |                                   |                                   |                                   |                                   |                                   |
|                     | B <sub>1</sub> | A <sub>1</sub> A <sub>2</sub>     | <b>A<sub>2</sub>A<sub>2</sub></b> |                                   |                                   |                                   |                                   |                                   |                                   |                                   |
|                     | C <sub>1</sub> | A <sub>1</sub> A <sub>3</sub>     | A <sub>2</sub> A <sub>3</sub>     | <b>A<sub>3</sub>A<sub>3</sub></b> |                                   |                                   |                                   |                                   |                                   |                                   |
| Method <sub>2</sub> | A <sub>2</sub> | <b>A<sub>1</sub>B<sub>1</sub></b> | A <sub>2</sub> B <sub>1</sub>     | A <sub>3</sub> B <sub>1</sub>     | <b>B<sub>1</sub>B<sub>1</sub></b> |                                   |                                   |                                   |                                   |                                   |
|                     | B <sub>2</sub> | A <sub>1</sub> B <sub>2</sub>     | <b>A<sub>2</sub>B<sub>2</sub></b> | A <sub>3</sub> B <sub>2</sub>     | B <sub>1</sub> B <sub>2</sub>     | <b>B<sub>2</sub>B<sub>2</sub></b> |                                   |                                   |                                   |                                   |
|                     | C <sub>2</sub> | A <sub>1</sub> B <sub>3</sub>     | A <sub>2</sub> B <sub>3</sub>     | <b>A<sub>3</sub>B<sub>3</sub></b> | B <sub>1</sub> B <sub>3</sub>     | B <sub>2</sub> B <sub>3</sub>     | <b>B<sub>3</sub>B<sub>3</sub></b> |                                   |                                   |                                   |
| Method <sub>3</sub> | A <sub>3</sub> | <b>A<sub>1</sub>C<sub>1</sub></b> | A <sub>2</sub> C <sub>1</sub>     | A <sub>3</sub> C <sub>1</sub>     | <b>B<sub>1</sub>C<sub>1</sub></b> | B <sub>2</sub> C <sub>1</sub>     | B <sub>3</sub> C <sub>1</sub>     | <b>C<sub>1</sub>C<sub>1</sub></b> |                                   |                                   |
|                     | B <sub>3</sub> | A <sub>1</sub> C <sub>2</sub>     | <b>A<sub>2</sub>C<sub>2</sub></b> | A <sub>3</sub> C <sub>2</sub>     | B <sub>1</sub> C <sub>2</sub>     | <b>B<sub>2</sub>C<sub>2</sub></b> | B <sub>3</sub> C <sub>2</sub>     | C <sub>1</sub> C <sub>2</sub>     | <b>C<sub>2</sub>C<sub>2</sub></b> |                                   |
|                     | C <sub>3</sub> | A <sub>1</sub> C <sub>3</sub>     | A <sub>2</sub> C <sub>3</sub>     | <b>A<sub>3</sub>C<sub>3</sub></b> | B <sub>1</sub> C <sub>3</sub>     | B <sub>2</sub> C <sub>3</sub>     | <b>B<sub>3</sub>C<sub>3</sub></b> | C <sub>1</sub> C <sub>3</sub>     | C <sub>2</sub> C <sub>3</sub>     | <b>C<sub>3</sub>C<sub>3</sub></b> |

Mono-trait Mono-method  
 Mono-trait Hetero-method (Convergent Validity)  
 Hetero-trait Hetero-method (Divergent validity)  
 Hetero-trait Mono-method (Method Variance)

1. For Convergent validity: MTHM correlations should be relatively large and significantly different from zero.
  2. For Discriminant validity: It involves three different criteria, which are as follows
    - a. First, any MTHM correlation must be significantly larger than any of the other correlations located on the same row and column.
    - b. Second, a variable should correlate higher with the same variable measured by different methods MTHM correlations, than with different variables measured by the same method HTMM correlations.
    - c. Third, the pattern of correlations between HTMM and HTHM correlation triangles must be the same.
- Failure to meet the criteria indicates that the measures are corrupted potentially by method bias.

**Conclusion**

Validity is crucial for ensuring accurate, credible, and meaningful research findings. It instills confidence in measurements and assessments, allowing researchers to draw accurate

conclusions and make informed decisions. Validity ensures that researchers measure what they intend to measure, enhancing the relevance of the results. It enhances the credibility of research by demonstrating the trustworthiness of the instruments. Understanding and ensuring validity is essential for producing high-quality valid research outcomes that inform decision-making and advance knowledge in various fields.

## References

- Aravamudhan, N.R. and Krishnaveni, R. 2015. Establishing and reporting content validity evidence of training and development capacity building scale (TDCBS). *Management: Journal of Contemporary Management Issues*, 20(1): 131-158.
- Bodie, G.D., Jones, S.M., Vickery, A.J., Hatcher, L. and Cannava, K. 2014. Examining the construct validity of enacted support: A multitrait-multimethod analysis of three perspectives for judging immediacy and listening behaviours. *Communication Monographs*, 81(4): 495-523.
- Campbell, D. T. and Fiske, D. W. 1959. Convergent and discriminant validation by the multi-trait multi-method matrix. *Psychological Bulletin*, 56(2): 81-105.
- Carmines, E. G. and Zeller, R. A. *Reliability and Validity Assessment*, California: Sage Publications, 2007, p 72.
- Cohen, J. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4): 213.
- Crocker, L. and Algina, J. *Introduction to classical and modern test theory*. New York: Wadsworth Publishing Co Inc, 2006, p 527.
- Cronbach, L. J. Test validation. In Thorndike, R. L. (Ed.). *Educational Measurement*, 2<sup>nd</sup> edition, USA: American Council on Education, 1971, pp. 443-507.
- Cronbach, L.J. and Meehl, P.E. 1955. Construct validity in psychological tests. *Psychological Bulletin*, 52(4): 281.
- Graham, J.C. and Garton, B.L. 2003. Certification measures: Are they predictive of secondary agriculture teacher performance? *Journal of Agricultural Education*, 44(3): 54-65.
- Gumus, M.M. and Kukul, V. 2023. Developing a digital competence scale for teachers: validity and reliability study. *Education and Information Technologies*, 28(3): 2747-2765.
- Handage, S. and Chander, M. 2021. Development of an instrument for measuring the student learning outcomes: A content validation process. *Indian Journal of Extension Education*, 57(3): 1-7.
- Ibiyemi, A., Mohd A. Y., Daud, M.N., Olanrele, S. and Jogunola, A. 2019. A content validity study of the test of valuers' support for capturing sustainability in the valuation process in Nigeria. *Pacific Rim Property Research Journal*, 25(3): 177-193.

- Jain, M. K. *Research Methodology and Statistical Techniques*, New Delhi: Shree Publishers & Distributors, 2005. p 290.
- Kelley, T. L. *Interpretation of Educational Measurements*, New York: World Book Company, 1927, p 392.
- Kerlinger, F. N. *Foundation of Behavioral Research*, New York: Holt, Rinehart and Winston, 2003, p 741.
- Lawshe, C.H. 1975. A quantitative approach to content validity. *Personnel Psychology*, 28(4): 563-575.
- Nunnally, J. C. *Psychometric Theory*, New York: McGraw-Hill, 1978, p 752.
- O'Leary, S.W. and Vokurka, R.J. 1998. The empirical assessment of construct validity. *Journal of Operations Management*, 16(4): 387-405.
- Patil, S. and Aditya. *Research Methodology in Social Sciences*, New Delhi: New India Publishing Agency, 2016, p 148.
- Pyo, J., Lee, W., Choi, E.Y., Jang, S.G. and Ock, M. 2023. Qualitative research in healthcare: necessity and characteristics. *Journal of preventive medicine and public health*, 56(1): 1-12.
- Rovinelli, R. J., & Hambleton, R. K. 1977. On the Use of Content Specialists in the Assessment of Criterion-Referenced Test Item Validity. *Dutch Journal of Educational Research*, 2: 49-60.
- Schmitt, N. and Stults, D.M. 1986. Methodology review: Analysis of multitrait-multimethod matrices. *Applied Psychological Measurement*, 10(1): 1-22.
- Schwab, S., Zurbriggen, C.L. and Venetz, M. 2020. Agreement among student, parent and teacher ratings of school inclusion: A multitrait-multimethod analysis. *Journal of School Psychology*, 82: 1-16.
- Shrotryia, V.K. and Dhanda, U. 2019. Content validity of assessment instrument for employee engagement. *Sage Open*, 9(1): 1-7.
- Singh, S.S., Singh, R.J., Devarani, L., Hemochandra, L. and Singh, R. 2023. Development of reliability and validity of social cohesiveness rating scale (SCRS). *Indian Journal of Extension Education*, 59(1): 139-141.
- Turner, Ronna, C., & Carlson, L. 2003. Indexes of Item-Objective Congruence for Multidimensional Items. *International Journal of Testing*, 3(2): 163-171.
- Wang, B., Rau, P.L.P. and Yuan, T. 2023. Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale. *Behaviour & information technology*, 42(9): 1324-1337.
- White, T.L. and McBurney, D.H. 2012. *Research methods*. 9<sup>th</sup> ed., USA: Cengage Learning, p 481.
- Yaghmaie, F. 2003. Content validity and its estimation. *Journal of medical education*, 3(1): 25-27.