Journal of Innovative Learning
Institute for Innovative Learning, Mahidol University

Vol.1 No.1, pages 13 - 18 Technical Article

Adapting the Index of Item-Objective Congruence to Items with Multiple Objectives

Parames LAOSINCHAI[a*]

[a] *Institute for Innovative Learning, Mahidol University, Thailand*
*Corresponding author: parames.lao@mahidol.edu

Received: February 6, 2023; Revised: March 6, 2024; Accepted: August 28, 2024

Abstract Finding the content validity of a test is a crucial step in assuring its quality. Enlisting the help of a team of content specialists to rate each item on the test is the simplest and most effective way to do so. Rovinelli and Hambleton (1977) adapted a procedure proposed by Hemphill and Westie (1950) and adjusted the latter's formula to compute the Index of Item-Objective Congruence (IIOC) which can be used for measuring content validity. This article explains the mathematical reasons behind the IIOC formula and simplifies its form so that the new form can be easily adapted to a test item with multiple objectives. In addition, this form can be used with any rating scale. Two extensions of the IIOC formula proposed by another team of authors are shown to be incompatible with each other and with the form derived in this article.

Keywords: IIOC, content validity, multiple objectives

Introduction

One of the major concerns when developing a test to measure a set of objectives is whether the test actually measures the intended objectives. A simple yet effective way to determine this is to enlist the help of a team of content specialists to rate each item on the test. After a test developer specifies the set of objectives to be measured and generates all the test items, it is up to a team of specialists to rate these items relative to these objectives. Although there usually are several objectives, far too many test developers restrict each specialist to rate each item relative to only one objective, the one that the test developers intend for the item to measure. This practice eliminates the specialist's freedom to choose the objective(s) that may be more closely related to the item than the one specified by the test developers (Turner and Carlson, 2003).

Provided that a test developer does not influence specialists' decisions, the next concern is to select an appropriate rating task. As pointed out by Rovinelli and Hambleton (1977), the rating task given to specialists should be simple, clear, non-tedious, and non-time-consuming. They described three procedures for collecting specialists' judgments: the Hemphill and Westie's (1950) categorizing procedure, the semantic differential rating procedure, and the matching procedure. The first procedure employed a scale with three familiar ratings: +1 if a specialist felt that an item measured an objective; 0 if a specialist was undecided whether an item measured an objective; and −1 if a specialist felt that an item did not measure an objective. As the names suggest, the second procedure utilized a semantic differential scale (Osgood et al., 1957) while the last did not involve any scale. Rovinelli and Hambleton conducted an empirical study of the three procedures and concluded that the Hemphill-Westie categorizing procedure was most suitable for collecting content specialists' judgments about whether an item measured an intended objective. It is this procedure which is the focus of this article.

Index of homogeneity of placement

Hemphill and Westie (1950) developed a series of scales to objectively describe group characteristics. Among them was the index of homogeneity of placement which was used for constructing a personality test. Once a test developer selected a set of personality dimensions and came up with test items, a team of psychological content specialists would judge whether each item applied to any specified dimensions. In this kind of personality test, it was typical for one item to apply to only one dimension. Thus, for an item to be selected, it had to clearly apply to one dimension and not to any others. Table 1 shows a template of the form in which each expert judge had to fill.

Having collected the judgments from all the experts, the index of homogeneity of placement for item k on dimension i, I_{ik}, can be calculated as follows:

$$I_{ik} = \frac{N\sum_{j=1}^{n} X_{ijk} - \sum_{l=1}^{N}\sum_{j=1}^{n} X_{ljk}}{2\cdot 2n(N-1) + \sum_{l=1}^{N}\sum_{j=1}^{n} X_{ljk} - \sum_{j=1}^{n} X_{ijk}}$$

where

N is the number of personality dimensions,

n is the number of content specialists, and

X_{ijk} is the rating of item k as a measure of dimension i by content specialist j.

To make sense of this formula, let us first look at example ratings of one item as shown in Table 2. The first term in the numerator results from substituting the ratings on the intended dimension (in black) for the ratings on other dimensions (in gray) and then summing all the ratings. The second term in the numerator (double summations) is just the summation of all the original ratings. Thus, the numerator represents the extent to which the ratings on the intended dimension exceed the ratings on other dimensions. The term $n(N-1)$ in the denominator represents the number of ratings on other dimensions while the leading factor 2 comes from the maximum difference between the two most extreme ratings (+1 and −1). Multiplied together, they represent the maximum magnitude of the numerator. Without other terms in the denominator, the quotient already makes sense as it represents the collective opinion of how well this item applies to the intended objective, normalized to always fall between −1 and 1. Other terms in the denominator were probably included by mistake.

Table 1 Template of the form for a content specialist with N personality dimensions and M test items. A value of +1 indicates that the item applies to the dimension; 0: indecisiveness; −1: the item does not apply to the dimension. Item 1 applies to dimension 1 only. Item 2 applies to dimension N and may or may not apply to dimension 2. Item M applies to dimensions 2 and N.

Item	Dimension 1	Dimension 2	…	Dimension *N*
1	+1	−1	…	−1
2	−1	0	…	+1
⋮	⋮	⋮		⋮
M	−1	+1	…	+1

Table 2 Example of ratings of an item on five personality dimensions by four judges. Dimension 3 (in black) is the intended one for this item.

Judge	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5
1	+1	−1	+1	−1	−1
2	0	0	+1	−1	−1
3	+1	0	+1	0	−1
4	0	−1	+1	−1	−1
Average	0.5	−0.5	1	−0.75	−1

Index of Item-Objective Congruence (IIOC)

Rovinelli and Hambleton (1977) were interested in assessing the content validity of items in a criterion-referenced test. They realized that one could adapt Hemphill and Westie's procedure simply by substituting "objective" for "personality dimension" and only the term $2n(N-1)$ remains in the denominator. They coined the term IIOC for the new index and rearranged the formula by eliminating the summation of the ratings on the intended objective from both terms in the numerator to obtain the formula

$$I_{ik} = \frac{(N-1)\sum_{j=1}^{n} X_{ijk} - \sum_{\substack{l=1 \\ l\neq i}}^{N} \sum_{j=1}^{n} X_{ljk}}{2n(N-1)}.$$

The quality of IIOC depends largely on the panel of judges. After collecting the ratings, one should always try to determine how much the judges agree about the ratings. Several methods to find the inter-rater reliability are available but Rovinelli and Hambleton recommended using Lu's (1971) coefficient of agreement for each objective separately. Hambleton et al. (1978) and Thorn and Deitz (1989) provided good explanations of IIOC and supported the use of Lu's coefficient of agreement.

Another characterization of IIOC

The first term in the numerator of IIOC formula now consists of $N-1$ copies of ratings on the intended objective k, which cancels out the same factor in the denominator. Each copy contains n ratings, one from each expert. Thus, when the sum is divided by n in the denominator, it yields the average of the ratings of item k on objective i. That is, the first term divided by $n(N-1)$ represents the average opinion of experts about item k on objective i.

The inner summation of the second term in the numerator contains n ratings of item k on one of the unintended objectives. Since there are $N-1$ such objectives, this second term contains $n(N-1)$ ratings. Thus, when the sum is divided by the same factor in the denominator, it yields the average of the ratings of item k on unintended objectives. That is, the second term divided by $n(N-1)$ represents the average opinion of experts about item k on unintended objectives.

Since $n(N-1)$ in the denominator is divided into both terms in the numerator, only the factor 2 remains. Again, it is the difference between the maximum and minimum ratings. Let m denote this difference, μ_{ik} denote the average opinion of experts about item k on objective i, and $\mu_{\bar{\imath}k}$ denote the average opinion of experts about item k on objectives other than i. The formula for IIOC now becomes

$$I_{ik} = \frac{\mu_{ik} - \mu_{\bar{\imath}k}}{m}. \tag{1}$$

Crocker and Algina (1986) gave another characterization of IIOC. Let μ_k denote the average opinion of experts about item k on all objectives. They stated that

$$I_{ik} = \frac{N(\mu_{ik} - \mu_k)}{2N-2}. \tag{2}$$

To see whether the two characterizations are equivalent, notice that

$$\mu_k = \frac{\mu_{ik} + (N-1)\mu_{\bar{\imath}k}}{N}. \tag{3}$$

That is, the overall average opinion is just the weighted average of the two average opinions, one on objective i and the other on objectives other than i. Substituting (3) into (2) and grouping the terms containing μ_{ik} yields

$$I_{ik} = \frac{N[(N-1)\mu_{ik} - (N-1)\mu_{\bar{\imath}k}]}{N(2N-2)}$$

which is the same as (1) after canceling $N(N-1)$ from both the numerator and the denominator and replacing 2 with m.

Adapting IIOC to a test item with multiple objectives

Not only can (1) be used for any rating scale, but it can also be adapted to a case where each test item may measure multiple objectives. The only thing we have to do is interpret i as representing a set of valid objectives and $\bar{\imath}$ as the complement of i. For any item k, the set of valid objectives i and IIOC can be found using the following steps:

1. Let $\bar{\imath}$ contain all the objectives and let IIOC be 0.
2. If all the objectives in $\bar{\imath}$ have the same average, stop.
3. Among the objectives in $\bar{\imath}$, move those (maybe only one) with the highest average to i.
4. Calculate the new IIOC according to (1).
5. If the new IIOC is greater than IIOC, set IIOC to the new value and go to step 2. Otherwise, reverse the most recent steps 3 and 4 and stop.

If the resulting IIOC is sufficiently high (see Hambleton et al., 1978; Rovinelli & Hambleton, 1977; and Thorn & Deitz, 1989 for guidelines), the set of objectives i should be taken as the objectives upon which the experts agree to be measured by item k. When all the objectives have the same average and μ_k is high enough, it can be taken as IIOC and item k has been judged as measuring all the objectives.

As an illustrative example, consider the ratings in Table 2 (substitute objective for dimension). The first objective to be moved to set i in step 3 is objective 3 which yields IIOC of

$$\frac{1-(0.5-0.5-0.75-1)/4}{2} = 0.71875.$$

The second objective to be moved to set i is objective 1 which results in IIOC of

$$\frac{(1+0.5)/2-(-0.5-0.75-1)/3}{2} = 0.75$$

which is greater than the preceding value. It is obvious that moving the next objective (objective 2) to set i would lower IIOC and the procedure would stop after reversing this move in step 5. Thus, item k measures objectives 1 and 3 with IIOC of 0.75. Notice that had the average opinion on objective 1 be any lower, the procedure would yield only objective 3 as the valid objective. Thus, experts' opinions are all that matters when deciding whether an item measures one or multiple objectives. As a result, IIOC always falls between 0 (instead of -1) and 1.

The procedure above employs only the average opinions. That is, once the averages have been calculated, one needs not consider each expert's opinion individually. In effect, the procedure finds the set of objectives i that maximizes IIOC for each item. In practice, one can almost always determine the set i without following the procedure. Let us use the example in Table 3 to demonstrate how to do this.

Experts unanimously agree that item 1 measures objectives 1 and 2 but does not measure objectives 3–5. So $\mu_{\{1,2\}1} = 1$, $\mu_{\{3,4,5\}1} = -1$, and $I_{\{1,2\}1} = (1-(-1))/2 = 1$. For item 2, experts tend to think that it measures objectives 2 and 3 but does not measure objectives 1, 4, and 5. Thus $\mu_{\{2,3\}2} = (0.75+0.5)/2 = 0.625$, $\mu_{\{1,4,5\}2} = (-1-0.5-0.75)/3 = -0.75$, and $I_{\{2,3\}2} = (0.625-(-0.75))/2 = 0.6875$. The average experts' opinions for items 3 and 4 are mirror images of each other. The question is whether the objective about which experts' opinions are inconclusive should be included in the set i. One can see that, for item 3, adding objective 4 to the set $i = \{5\}$ would lower μ_{ik} by 0.5 while adding it to the set $\bar{\imath} = \{1,2,3\}$ would increase $\mu_{\bar{\imath}k}$ by only 0.25. That is, the latter choice would maximize IIOC. The opposite is true for item 4. The procedure tends to put an inconclusive item on the side with more objectives.

Table 3 Average experts' opinions and Index of Item-Objective Congruence when some items measure multiple objectives. The symbol * indicates the objective(s) measured by each item.

Item	Average experts' opinions on objective					IIOC
	1	2	3	4	5	
1	1*	1*	−1	−1	−1	1
2	−1	0.75*	0.5*	−0.5	−0.75	0.6875
3	−1	−1	−1	0	1*	0.875
4	−1	0*	1*	1*	1*	0.875

Discussion

Turner and Carlson (2003) proposed another formula for IIOC that could be applied to multidimensional items. The formula was

$$I_{ik} = \frac{N\mu_{ik} - (N - |i|)\mu_{\bar{\imath}k}}{2N - |i|} \tag{4}$$

where $|i|$ denoted the number of objectives in set i. In special cases where $\mu_{ik} = -\mu_{\bar{\imath}k}$, I_{ik} would equal to μ_{ik} which agrees with IIOC computed by using (1). However, if $\mu_{ik} \neq -\mu_{\bar{\imath}k}$, the two formulas would yield different results. For the ratings in Table 2, if set i contains only objective 3, (4) would yield IIOC of

$$\frac{5\cdot 1 - (5-1)(0.5 - 0.5 - 0.75 - 1)/4}{2\cdot 5 - 1} = 0.75$$

instead of 0.71875, which would lead to the conclusion that we may take only objective 3 (instead of objectives 1 and 3) to be measured by this item.

Turner and Carlson also referred to Turner et al. (2002) for a SAS/IML® macro that could be used for computing (4). However, the latter article gave the formula for IIOC as

$$I_{ik} = \frac{(N + 2|i| - 2)\sum_{j=1}^{n} X_{ijk} - |i|\sum_{l=1}^{N}\sum_{j=1}^{n} X_{ljk}}{2n(N-1)|i|}. \tag{5}$$

When $|i| = 1$, this reduces to the original IIOC formula. Hence, it does not agree with (4). On the other hand, when $\mu_{ik} = -\mu_{\bar{\imath}k}$, I_{ik} would again equal to μ_{ik} which equals to the true IIOC. Thus, we cannot test this formula with the ratings in Table 2. Let us change all the ratings for objective 1 to be +1 and set i contains objectives 1 and 3. IIOC computed by using (1) is

$$\frac{(1+1)/2 - (-0.5 - 0.75 - 1)/3}{2} = 0.875$$

while that computed by using (5) is

$$\frac{(5 + 2\cdot 2 - 2)\cdot 8 - 2\cdot(-1)}{2\cdot 4\cdot(5-1)\cdot 2} = 0.90625.$$

Thus, (5) is not equivalent to (1).

In conclusion, among the three formulas for IIOC (formulas (1), (4), and (5)) that may be adaptable to test items with multiple objectives, the one derived in this article (formula (1)) is the only sensible adaptation. Furthermore, it is the most elegant and can be used with any rating scale. The authors of the other formulas never explained why their formulas should take those forms.

The proposed formula and procedure for IIOC allow experts to give their opinions without the influence of the test developer. Besides, both are easily implemented. All a test developer has to do is prepare the lists of both the items and the objectives and a table similar to Table 1. Once the ratings are obtained, IIOC can be calculated manually as illustrated in the previous section or with the help of an electronic spreadsheet.

References

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt Brace Jovanovich, Inc.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, *48*(1), 1–47. https://doi.org/10.3102/00346543048001001

Hemphill, J. K., & Westie, C. M. (1950). The measurement of group dimensions. *The Journal of Psychology*, *29*(2), 325–342. https://doi.org/10.1080/00223980.1950.9916035

Lu, K. H. (1971). A Measure of agreement among subjective judgments. *Educational and Psychological Measurement*, *31*(1), 75–84. https://doi.org/10.1177/001316447103100105

Osgood, C. E., Sucic, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. University of Illinois Press.

Rovinelli, R. J., & Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Dutch Journal for Educational Research*, *2*(1), 49–60.

Thorn, D. W., & Deitz, J. C. (1989). Examining content validity through the use of content experts. *The Occupational Therapy Journal of Research*, *9*(6), 334–346. https://doi.org/10.1177/153944928900900602

Turner, R. C., & Carlson, L. (2003). Indexes of item-objective congruence for multidimensional items. *International Journal of Testing*, *3*(2), 163–171. https://doi.org/10.1207/S15327574IJT0302_5

Turner, R. C., Mulvenon, S. W., Thomas, S. P., & Balkin, R. S. (2002). Computing indices of item congruence for test development validity assessments. In *Proceedings of the Twenty-Seventh Annual SAS® Users Group International Conference* (pp. 255.1–5). SAS Institute Inc. https://support.sas.com/resources/papers/proceedings/proceedings/sugi27/p255-27.pdf