Item validity vs. item discrimination index: a redundancy?

R L Panjaitan1*, R Irawati1, A Sujana1, N Hanifah1 and D Djuanda1

1Pendidikan Guru Sekolah Dasar, Kampus Sumedang, Universitas Pendidikan Indonesia, Jalan Mayor Abdurrahman 211, Sumedang 45322, Indonesia

*Corresponding Author: regina@upi.edu

Abstract. In several literatures about evaluation and test analysis, it is common to find that there are calculations of item validity as well as item discrimination index (D) with different formula for each. Meanwhile, other resources said that item discrimination index could be obtained by calculating the correlation between the testee’s score in a particular item and the testee’s score on the overall test, which is actually the same concept as item validity. Some research reports, especially undergraduate theses tend to include both item validity and item discrimination index in the instrument analysis. It seems that these concepts might overlap for both reflect the test quality on measuring the examinees’ ability. In this paper, examples of some results of data processing on item validity and item discrimination index were compared. It would be discussed whether item validity and item discrimination index can be represented by one of them only or it should be better to present both calculations for simple test analysis, especially in undergraduate theses where test analyses were included.

1. Introduction

In educational framework related to evaluation, both item validity (IV) and item discrimination index (D) are commonly found as part of item analysis. Many undergraduate theses (i.e. 44 out of 45 students’ theses of Elementary School Teacher Education with mathematics field-of-interest at Universitas Pendidikan Indonesia Sumedang campus in 2017) applying quantitative method in education included IV and D calculations as some considerations in judging whether an item is acceptable to be used in a test or not.

Nonetheless, there are various views associated with IV and D. In general, validity is considered as the capability of an instrument to assess what the tester would like to assess [1-2]. Meanwhile, item discrimination index (D) is the indicator whether a question could distinguish between good and underachieving students [3-4]. Furthermore, item validity is the correlation between the testees’ on single item score and their total score [5-6], that can be calculated with equation (1), which is actually a Pearson product-moment correlation defined as

\[ r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2\sum(Y-\bar{Y})^2}} \]  

where \(X\) and \(Y\) are the testees’ score on the particular measured item and the testees’ total score for all item, respectively [5-6].

It is also said that item validity could be calculated with point-biserial coefficient calculation [5], while other resources mentioned that point-biserial coefficient is used to compute item discrimination index [7-8]. In addition, for essay questions, item discrimination index could be computed with...
Kelly’s approach with the same interpretation as point-biserial calculation [8]. It seems that \( IV \) and \( D \) are the same concept in item analysis, some authors gave different (or same) formula to calculate both, though.

As mentioned above, item discrimination index \( (D) \) has various ways to compute. However, \( D \) discussed in this paper was \( D \) calculation which was often used in undergraduate educational theses involving essay question item analysis. For this \( D \) calculation, when the testees were first grouped into upper group and lower group with the same number of testees, then the equation is found as

\[
D = \frac{\bar{x}_U - \bar{x}_L}{\text{max. score}}
\]  

(2)

where \( \bar{x}_U \) is the mean of score obtained by the upper group, \( \bar{x}_L \) is the mean of score obtained by the lower group and \( \text{max. score} \) is the maximum score of the item [9].

In this paper, the authors compared the result of equation (1) and equation (2) applied in the same data-set and examined whether those equations actually can be represented by only one equation or not.

2. Methods

This work was carried out with literature research methodology. Literature research methodology is to explore sources of reading for pointing out important features of information [10]. Data were collected from three theses of undergraduate students of elementary school teacher education study program at Universitas Pendidikan Indonesia Sumedang campus. Content analysis, which was the approach to examine reading sources to acquire finding [11], was applied in this study. Thus, this findings were scrutinized and then linked one to another to result literature-based discovery (LRD), which is original knowledge developed from connecting two or more literature concepts [12-13]. LRD in this study would be the information whether item validity \( (IV) \) and item discrimination index \( (D) \) can be depicted by only one of them or both \( IV \) and \( D \) should appear in any item analysis in experimental-design research.

3. Result and Discussion

All of the data-sets in this discussion were taken from undergraduate educational theses applying experimental-design research [14-16]. Data-set A, B and C were obtained from reference [14-16] respectively. Item validity and item discrimination index in Table 1, Table 2, Table 3, Figure 1, Figure 2 dan Figure 3 were the results from raw data processing with equation (1) and equation (2).

### Table 1. Item Validity and Discrimination Index for data-set A [14]

| Question Number | Item Validity (IV) | Discrimination Index (D) |
|-----------------|-------------------|-------------------------|
| 1               | 0.113             | 0                       |
| 2               | 0.578             | 0.33                    |
| 3               | 0.756             | 0.59                    |
| 4               | 0.549             | 0.33                    |
| 5               | 0.648             | 0.43                    |
| 6               | 0.205             | 0                       |
| 7               | 0.576             | 0.38                    |
| 8               | 0.695             | 0.67                    |
| 9               | 0.593             | 0.3                     |
| 10              | 0.548             | 0.23                    |
| 11              | 0.598             | 0.47                    |
| 12              | 0.616             | 0.47                    |
| 13              | 0.595             | 0.33                    |
| 14              | 0.504             | 0.3                     |
| 15              | 0.792             | 0.47                    |
| 16              | 0.515             | 0.37                    |
Table 2. Item Validity and Discrimination Index for data-set B [15]

| Question Number | Item Validity (IV) | Discrimination Index (D) |
|-----------------|-------------------|--------------------------|
| 1               | 0.75              | 0.36                     |
| 2               | 0.81              | 0.18                     |
| 3               | 0.55              | 0.39                     |
| 4               | 0.77              | 0.7                      |
| 5               | 0.73              | 0.32                     |
| 6               | 0.65              | 0.22                     |
| 7               | 0.39              | 0.29                     |
| 8               | 0.36              | 0.03                     |

Table 3. Item Validity and Discrimination Index for data-set C [16]

| Question Number | Item Validity (IV) | Discrimination Index (D) |
|-----------------|-------------------|--------------------------|
| 1               | 0.087             | 0.025                    |
| 2               | 0.537             | 0.333                    |
| 3               | 0.639             | 0.208                    |
| 4               | 0.599             | 0.437                    |
| 5               | 0.319             | 0.437                    |
| 6               | 0.429             | 0.541                    |
| 7               | 0.576             | 0.275                    |
| 8               | 0.507             | 0.5                      |
| 9               | 0.426             | 0.25                     |
| 10              | 0.654             | 0.125                    |
| 11              | 0.403             | 0.145                    |
| 12              | 0.633             | 0.208                    |
| 13              | 0.786             | 0.5                      |
| 14              | 0.541             | 0.312                    |
| 15              | 0.628             | 0.25                     |
| 16              | 0.685             | 0.395                    |

Data-sets in Table 1, Table 2 and Table 3 were transformed to be graphs that can be seen in Figure 2, Figure 2 and Figure 3.

![Figure 1](image-url)  
Figure 1. Item Validity and Discrimination Index for data-set A.
From the data in table 1, table 2 and table 3, the attained Pearson correlations ($r$) between $IV$ and $D$ are 0.898, 0.464 and 0.312 respectively. Meanwhile, the Pearson’s correlation critical value for $N = 16$ (data-set A and C) is 2.12 (level of significance 0.05) and for $N = 8$ (data-set B) is 2.306 (level of significance 0.05). The observed $t$ for data-sets A, B and C are 7.643, 1.284 and 1.227. Thus, from the comparison of observed $t$ for each data-set to their Pearson’s correlations critical values, it can be inferred that there is significant correlation between $IV$ and $D$ for data-set A. Nevertheless, there is no significant correlation between $IV$ and $D$ for data-set B and C.

The significant correlation between $IV$ and $D$ can be clearly seen from figure 1. However, the graph shows that for question number 1 and 6, when $IV$ is 0.113 and 0.205, $D$ values for both are zero. It can be observed that $IV$ and $D$ for data-set A have the same patterns, and $IV$ was always higher than $D$.

In addition, there is no significant correlation between $IV$ and $D$ for both data-set B and data-set C. In spite of those correlation insignificance, it can be easily perceived that in data-set A and B, $IV$ was always higher than $D$. Data-set C also acts almost similarly to data-set A and B, in the way of $IV$ was generally higher than $D$ for almost all datapoints, the datapoints of question 5 and question 6 were the outliers, though.

As it is said before, conceptually for some views $IV$ and $D$ are alike. But if they are referred from equation (1) and equation (2), $IV$ and $D$ calculations resulted quite much different values. However, although not all of the datapairs of $IV$ and $D$ have significant correlations, three data-sets examined in this paper show that $IV$ and $D$ generally has the same patterns and $IV$ were almost always higher than $D$. This leads to the consequences that $IV$ calculation itself cannot be assumed as the only adequate indicator when judging the quality of item in evaluation is needed. It is often found that the tester calculated both $IV$ and $D$, but he/she only relied on $IV$ for judging which question item that should be used and which item that should be revised or omitted. From figure 1, figure 2 and figure 3 it can be noticed that more sensitive results were derived from $D$ calculation instead of $IV$. The smaller values of $D$ (compared to $IV$) give more assurance to the tester when judging an item question. For example, when $D = 0$ for question number 6 data set A, $IV = 0.205$. If the tester only relies on $IV$ for judging an item without considering $D$, question number 6 could be fairly acceptable (which is actually not a good judging decision). But if the tester only relies on $D$ (which is zero), the tester would already realized that that item cannot function perfectly in discriminating good and poor students, and the judging can be done even without $IV$ calculation.

4. Conclusion
There is low significance of correlation between $IV$ and $D$ in two data-sets (from three data-sets), so it cannot be concluded that $IV$ (calculated with equation (1)) and $D$ (calculated with equation (2)) has redundancy effect.

However, it can be seen from the discussion that in almost cases, $IV$ has higher index value than $D$. Thus, when judging an item question, generally $D$ is more sensitive than $IV$. A researcher cannot judge
a quality of item from IV calculation only, without taking D into consideration. Researchers, especially undergraduate students, are expected to be more aware for not relying their judging merely on IV calculation, because in many cases, although an item has good IV, it could have poor D or even zero in D value. Zero in D value means that the item cannot distinguish the ability of the testees. It depends on the researchers whether they need to compute both IV and D. However, if there are some cases that the researcher has to rely on only one choice between IV and D calculation in judging an item question, D calculation would be generally better to consider than IV calculation.

This kind of work can be continued with considering more data-sets and different formula for IV and D. More data-sets involved would obtain more robust results and interpretations.

References
[1] Colton D and Covert R W 2007 Designing and Constructing Instruments for Social Research and Evaluation (San Francisco: Jossey-Bass)
[2] Dimitrov D M 2012 Statistical Methods for Validation of Assessment Scale Data in Counseling and Related Fields (Alexandria: American Counseling Association)
[3] Khan G-A N, Ishrat N and Khan A Q 2015 Int. J. Res. Med. Sci. 3 178
[4] Taib F and Yusoff M S B 2014 J. of Taibah University Med. Sci. 9 110
[5] Arikunto S 2013 Dasar-dasar Evaluasi Pendidikan edisi 2 (Jakarta: Bumi Aksara)
[6] Jihad A and Haris A 2013 Evaluasi Pembelajaran (Yogyakarta: Multi Pressindo)
[7] Attali Y and Fraenkel T 2000 J. Educ. Meas. 37 77
[8] Wright R J 2008 Educational Assessment Tests and Measurements in the Age of Accountability (Thousand Oaks: SAGE Publications, Inc)
[9] Arifin Z 2009 Evaluasi Pembelajaran, Prinsip Teknik Prosedur (Bandung: PT Remaja Rosdakarya)
[10] Lin G 2009 Int. Educ. Stud. 2 179
[11] White M D and Marsh E E 2006 Library Trends 55 22
[12] Kostoff R N 2008 Technol. Forecast. Soc. Change 75 165
[13] Kostoff R N 2008 Technol. Forecast. Soc. Change 75 186
[14] Anugrah R 2017 The Effect of Learning Cycle on Student’s Science Process Skills on the Magnetic Force Learning [unpublished undergraduate thesis] (Sumedang: Universitas Pendidikan Indonesia)
[15] Mariyana G 2017 The Influence of Learning-Start-with-Questions to the Critical Thinking Skill of Elementary School on the Material Mirror [unpublished undergraduate thesis] (Sumedang: Universitas Pendidikan Indonesia)
[16] Kaffah S 2017 The Effect of Group Investigation Method to the Students’ Creative Thinking Ability on the Topic of Natural Disaster [unpublished undergraduate thesis] (Sumedang: Universitas Pendidikan Indonesia)