RedFame

Journal of Education and Training Studies
Vol. 6, No. 7; July 2018
ISSN 2324-805X E-ISSN 2324-8068
Published by Redfame Publishing
URL: http://jets.redfame.com

Using Classroom Observation Scores Instead of Test Scores as Criterion in the Estimation of Discrimination Index

Esin Bağcan Büyükturan[1], Ayşe Şireci[2]

[1]Education Faculty, Abant İzzet Baysal University, Bolu, Turkey

[2]Ömer Nasuhi Bilmen Secondary School, Ministiry of Education, Şanlıurfa, Turkey

Correspondence: Esin Bağcan Büyükturan, Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Oda No: 220 Gölköy Kampüsü Bolu, Turkey.

Received: April 9, 2018 Accepted: April 28, 2018 Online Published: May 25, 2018

doi:10.11114/jets.v6i7.3191 URL: https://doi.org/10.11114/jets.v6i7.3191

Abstract

Item discrimination index, which indicates the ability of the item to distinguish whether or not the individuals have acquired the qualities that are evaluated, is basically a validity measure and it is estimated by examining the fit between item score and the test score. Based on the definition of item discrimination index, classroom observation scores were used in this study instead of test scores as the indication of having the tested quality.

In the framework of the study, a 25-item multiple-choice test prepared in the context of 8th grade Mathematics Unit "Multipliers and Multiples" was administered to a total of 109 8th graders (44 females, 65 males) studying in 4 separate classrooms of Ömer Nasuhi Bilmen Secondary School in Şanlıurfa Province. Furthermore, these students' Mathematics teachers were asked to observe and score students during the unit and the obtained observation scores were used as external criterion in estimating the discrimination index. By using this criterion, fit values estimated with the help of *upper* and *lower groups* consisting of *27*% from the extremes of the criterion score distribution and biserial correlation were compared with the traditional conditions where test scores were utilized. It was found that item discrimination indices based on classroom observations were higher than those based on test scores in both of the discrimination indices estimated via upper-lower 27% groups and biserial correlation. This finding was discussed to be related to the fact that while classroom observation scores were an external validity criterion, test scores were composed of items whose discrimination values were calculated. The finding also demonstrated that classroom observation scores were more rigid and eliminative than test scores in terms of discrimination.

Keywords: item discrimination index, test score, item score, classroom observation score

1. Introduction

Based on certain assumptions to solve basic measurement problems, classical test theory relies on estimation by using observed test scores. In this theory, the most basic parameters used in the process of item discrimination are item difficulty index and item discrimination index (Baykul, 2000; Crocker and Algina, 1986).

Item difficulty index is defined as the percentage of answering an item accurately and item becomes easier when this value is high. Item discrimination index is the power of the item to distinguish between individuals with or without the tested qualities, or in other words, the individuals who have or have not acquired the desired quality. Individuals' test scores are taken as the criterion in item discrimination index estimation to distinguish between the individuals with and without accurate answers, with and without the desired quality and the lower and upper groups. In the context of this criterion, the definition also includes the correlation between item scores and test scores and the power of distinguishing between individuals with and without the measured qualities as a whole. (Baykul, 2000; Kilmen, 2014; Demars, 2010).

The most widely used method for estimating item discrimination index is based on upper and lower group, however, this method is criticized for ruling out a significant part of the group (Crocker and Algina, 1986; Baykul, 2000; Kilmen, 2014). Apart from this method, methods based on the correlation between item score and test score are used as well. Both methods accept the score obtained from the test as a whole as a criterion for estimation.

The facts that formulas used in item discrimination index estimation take test scores as a criterion and that item scores which comply with the test score are considered as discriminators are based on the assumption that the test -which the

item belongs to- is accepted as a valid criterion. Since there is not sufficient evidence about the validity of a teacher-made test, alternatives can be developed to accept test scores as the only criterion for discrimination estimation.

Teachers' observation of students and evaluation of student performance during the process is as important as the tests composed of a llimited number of items which sample the topics and performed in a limited time frame (Stiggins and Bridgeford, 1985; Anderson, 1987; Baki and Birgin, 2002) As a matter of fact, since teachers' daily classroom observations provide opportunities for direct, unmediated and first-hand observation, they constitute the main elements for assessing student achievement (Salmon and Cox, 1981; Herman and Dorr, 1983; Airasian, 1979). Although test scores do not reflect student performance in its entirety and do not fully reveal student knowledge, they are preferred over performance-based classroom assessments due to their consistency and accountability. Teachers' in-class assessment is an informal activity based on asking questions, observing activities and monitoring task completion and is expected to have low level of consistency (Gipps, 1994). This indicates that paper-and-pencil tests are more reliable than teachers' classroom observations, but it does not change the fact that the teacher assessments based on classroom observations are more valid since classroom assessments are carried out in a wider spectrum (Parkes & Maughan, 2009).

Considering the scores obtained through teacher observations and assessments as an external validity criterion in addition to test scores in discrimination index estimation will strengthen validity evidence.

Validity estimate with respect to external criterion is based on calculating the correlation coefficient between two series of scores obtained from the same sample group. This correlation coefficient is a measure of the covariance of two series and the validity study conducted with this method is called convergent validity (Kağıtçıbaşı, 1976; Arıcı, 1992; Baykul, 1996; Turgut, 1983).

When it comes to validity, it is common to check for concordance with an external criterion. Scale development studies are the most typical examples to this. Concordance between a developed scale and an existing one is regarded as evidence of its validity.

Based on these practices; taking into account the fact that item discrimination index is also an evidence of concordance; it may be possible to check for concordance solely with an external criterion or by using the external criterion together with the criterion that utilizes the score obtained from the whole test in which the item used in item discrimination index estimation is included. This study aimed to compare the use of classroom observation and assessment scores as external criteria and the use of traditional test scores in estimating item discrimination index. For this purpose, significance of differences between mean discrimination indexes calculated based on upper-lower 27% brackets according to criteria (such as direction and significance of the relationship between classroom observation scores and test scores, test scores and classroom observation scores) were examined as well as the significance of differences between mean discrimination indexes calculated with the help of biserial correlation method according to the same criteria. Additionally, correlations related to increases in discrimination indexes that were obtained according to both criteria were investigated.

2. Method

2.1 Model

This study was conducted as a basic research to investigate the alterations in discrimination index in cases where the criterion variable was changed.

2.2 Participants

The study group was composed of a total of 109 8^{th} graders (44 females, 65 males) studying in 4 separate classrooms of Ömer Nasuhi Bilmen Secondary School in Şanlıurfa Province. Mathematics was taught by the same teacher in all participating classrooms.

2.3 Measurement Instrument

Achievement test for Mathematics lesson "Factors and Multipliers" Unit, used as the measurement tool in the study, was prepared by the teacher who taught Mathematics in all participating classrooms. In order to ensure content validity, a Table of Specifications was created which included the learning outcomes in the row and cognitive taxonomic level in the column. The test included 27 multiple choice items at first but the items were reviewed by two experts (one Mathematics teacher and one assessment and evaluation expert) in terms of content representation, conformity to multiple choice test preparation criteria and scientific accuracy and 2 problematic items were excluded from the test. Also, 3 items were revised based on suggestions to finalize the test. The final test included 25 multiple choice items.

The learning outcomes of the unit represented by the scope of the measurement instrument were assessed by the teacher through classroom observations and assessments and scored the acquisitions out of 100. The participating teacher obtained the classroom observation score in this manner.

2.4 Data Analysis

Data were analyzed at .05 level of significance and parametric statistical techniques were used when the normal distribution was satisfied. The normality assumption was investigated by Kolmogorov-Smirnov test. t-test was used to test the significance between the means and Pearson product-moment correlation coefficient technique was utilized to explore the correlations related to the increase of both variables. Fisher Exact test was used to investigate the relationship between categorical variables

3. Results

Table 1 presents the descriptive statistics for classroom observation scores and multiple choice test scores.

Table 1. Descriptive statistics for classroom observation scores and multiple choice test scores

	N	Minimum	Maximum	Mean	Std. Deviation
Classroom observation score (out of 100)	109	20	100	58.31	24.01
Test score (out of 100)	109	12	80	45.58	15.38

Table 1 demonstrates that classroom observation scores changed between 20 and 100 with a mean of 58,31; test scores varied between 12 and 80 with a mean of 45,58. t-test was used to compare the significance of the difference between the means and the results are provided in Table 2.

Table 2. Results of the t-test conducted to compare the means of classroom observation scores and test scores

Measurement	N	$\overline{X}$	S	df	t	p
Classroom observation score	109	58.31	24.01	108	6.17	0
Test score	109	45.58	15.38			

According to Table 2, t-test results demonstrate that means of classroom observation scores were significantly higher than the means of multiple choice test scores (t=6,17, p<.05).

Correlation between the increase in classroom observation scores and multiple choice test scores were examined and results are summarized in Table 3.

Table 3. Pearson Correlation Table for the correlation between classroom observation scores and multiple choice test scores

	Classroom observation score	Multiple choice test score
Classroom observation score		
r	1.00	0.474
p		0
N	109	109
Multiple choice test score		
r	0.474	1
p	0	
N	109	109

According to Table 3, a moderate, positive and meaningful relationship was identified between classroom observation cores and multiple choice test scores (r=0.474; p<.05). This finding indicates that scores provided by teacher observations and the test increase in correlation.

Table 4 displays the discrimination indexes for both measurements obtained by using multiple choice test items in discrimination index estimation as a criterion based on lower-upper 27% segment.

Table 4. Item discrimination indexes calculated according to multiple choice test scores and classroom observation scores criteria based on lower-upper 27% segment

	Multiple choice test score criterion	Classroom observation score criterion
Item 1	0.21	0.21
Item 2	0.31	0.24
Item 3	0.48	0.38
Item 4	0.38	0.21
Item 5	0.52	0.38
Item 6	0.62	0.34
Item 7	0.48	0.41
Item 8	0.21	**0.07**
Item 9	0.69	0.38
Item 10	0.59	0.24
Item 11	0.48	0.24
Item 12	0.21	**-0.07**
Item 13	0.31	**-0.07**
Item 14	0.41	**0.07**
Item 15	0.45	**0.14**
Item 16	0.48	0.38
Item 17	**0.14**	**-0.14**
Item 18	0.55	0.24
Item 19	0.38	**0.10**
Item 20	0.41	**0.17**
Item 21	0.38	0.28
Item 22	0.41	0.41
Item 23	**0.03**	**0**
Item 24	**-0.07**	**-0.10**
Item 25	0.41	0.34

Table 4 shows the discrimination indexes that are below and over the 0.20 critical value. Tekin (2000) accepted the discrimination of items with values below 0.20 as weak and indicated that they should be excluded from the test if they could not be revised. In this case, the expression "it would be erroneous to use the item in the test without revision" was valid only for 3 items out of 25 when multiple choice test scores were taken as the criterion in discrimination index estimation, and for 10 items when classroom observation scores were taken as the criterion in discrimination index estimation. The fact that more items were acceptable when test scores were taken as the criterion may be related to obtaining these scores from the test in its entirety. Teacher's classroom observations cores may be used as a more eliminative external criterion for researchers who seek to develop higher quality items.

Fisher Exact test was used to determine whether item discrimination index values below or over .20 were related to criterion and results are provided in Table 5.

Table 5. Fisher test results for the comparison of numbers of items with discrimination index below or over .20 according to criterion used in item discrimination index estimation calculated by using lower-upper 27% group method

	Multiple choice test scores criterion	Classroom observation scores criterion
Number of items with discrimination index below .20	3	10
Number of items with discrimination index over .20	22	15

Fisher=141.572; p=0

According to Table 5, in item discrimination index estimation based on lower-upper 27% group method; number of items below or over the item discrimination index critical value 0.20 significantly changed when the criterion changed; in other words, whether the discrimination index was below or over the critical value depended on the criterion (Fisher=141.572; $p<.05$).

Mean item discrimination index values obtained by using lower-upper 27% method when both measurement results were used as criteria were compared and results are provided in Table 6.

Table 6. t-test for the significance of differences between mean item discrimination index values obtained by using lower-upper 27% method when classroom observation scores and multiple choice test scores were accepted as criteria.

Discrimination index criterion	N	$\overline{X}$	S	df	t	p
Classroom observation score	25	0.194	0.121	24	7.609	0
Multiple choice test score	25	0.379				

According to Table 6, there was a statistically significant difference between mean item discrimination index values obtained by using classroom observation scores and multiple choice test scores as criteria and this difference was in favor of multiple choice test scores (t=7,609; p<.05). The fact that multiple choice test scores had a higher correlation with item scores, being a part of the test themselves, was expected. However, when the mode of examining the learning outcomes changed while distinguishing the students with and without acquisitions (i.e. when classroom observation scores were used as external criterion for this study); artificial similarities between criterion and item score will be removed. In this sense, while classroom observation scores had lower concordance with item scores, they were more realistic in distinguishing students with and without desired acquisitions.

Correlations for the increase in item discrimination indexes for these two criteria were explored and results are provided in Table 7.

Table 7. Relationship between item discrimination indexes estimated according to upper-power 27% method by using classroom observation scores and multiple choice test scores as criteria

	Classroom observation score	Multiple choice test score
Classroom observation score		
r	1.00	0.763
p		0
N	25	25
Multiple choice test score		
r	0.763	1
p	0	
N	25	25

Based on Table 7, a high level, positive and significant relationship existed between discrimination indexes estimated by taking multiple choice test scores and classroom observation scores as criteria according to upper-lower 27% method (r=0.763; p<.05). This finding indicated that both criteria listed the items similar to their discriminatory values.

Discrimination indexes were also estimated via another method used in item discrimination index estimation, biserial correlation method. In other words; biserial correlation coefficients between criterion score and item score were obtained and provided in Table 8.

Table 8. Item discrimination indexes estimated according to biserial correlation method

	Relationship with multiple choice test score		Relationship with classroom observation score	
	r	**p**	**r**	**p**
Item1	0,281*	0,003	0,251*	0,009
Item 2	0,249*	0,009	0,147	0,128
Item 3	0,412*	0	0,322*	0,001
Item 4	0,242*	0,011	0,125	0,197
Item 5	0,447*	0	0,360*	0
Item 6	0,460*	0	0,293	0,002
Item 7	0,436*	0	0,334*	0
Item 8	0,196*	0,041	0,072	0,460
Item 9	0,518*	0	0,203*	0,034
Item 10	0,459*	0	0,180	0,060
Item 11	0,429*	0	0,166	0,084
Item 12	0,174	0,07	-0,103	0,288
Item 13	0,249*	0,009	-0,092	0,346
Item 14	0,345*	0	0,050	0,606
Item 15	0,338*	0	0,098	0,310
Item 16	0,423*	0	0,292*	0,002
Item 17	0,148	0,125	-0,115	0,232
Item 18	0,476*	0	0,234*	0,014
Item 19	0,407*	0	0,093	0,337
Item 20	0,362*	0	0,186	0,053
Item 21	0,393*	0	0,323*	0,001
Item 22	0,419*	0	0,327*	0,001
Item 23	0,029	0,764	-0,077	0,426
Item 24	-0,088	0,365	-0,099	0,304
Item25	0,409	0	0,286*	0,003

While 20 item discrimination indexes calculated according to biserial correlation method by taking test score as the criterion were significant, only 10 item discrimination indexes calculated based on same method by taking classroom observation scores as the criterion were found significant. Fisher Exact test was utilized to test the significance of the difference related to criterion-based change in the number of items with and without significant correlation and the results are provided in Table 9.

Table 9. Fisher Test results for the comparison of number of items that were significant and insignificant according to the criterion used in estimating the discrimination indexes calculated by using biserial correlation method

	Criterion Multiple choice test scores	**Criterion Classroom observation scores**
Number of items with significant correlation	3	10
Number of items without significant correlation	22	15

Fisher=144.628; p=0

According to Table 9, in item discrimination index estimation with biserial correlation method; significance of item discrimination index meaningfully changed when criterion was changed. In other words; significance of biserial correlation depended on the criterion (Fisher=144.628; $p<.05$).

More and higher level correlations between test scores and items were expected considering the fact that items were a part of the test in question. However, classroom observation scores are completely an external criterion and therefore fewer significant correlations were detected between item scores and classroom observation scores. In this case, just as in upper-lower 27% method, classroom observation scores present a more difficult criterion for discrimination in item discrimination index estimation according to biserial correlation method while it was easier for items to be regarded as discriminatory when estimation was done according to test scores.

4. Discussion

Item discrimination index is an indicator of item validity. When it comes to validity, it is common to check concordance with an external criterion. Scale development studies are the most typical examples in this regard. For instance, in Demir's (2011) validity and reliability study of the Turkish version of Functions of Identity Scale, Psychological Well-being Scale was accepted as an external criterion and Kızılkaya ve Aşkar (2009) investigated students'

mathematics achievement scores as the external criterion when they developed Reflective Thinking Skill Scale towards Problem Solving. In this study, classroom observation scores were used as the external criterion for item discrimination.

Study results suggest that regardless of the method used in item discrimination index estimation, estimation based on item and test concordance provided higher concordance values compared to estimation based on classroom observation scores, regarded as the external criterion. While investigating the concordance between the item and the test, it should be kept in mind that items were a part of the test and therefore they might have artificially increased concordance values. This finding can be interpreted in a different manner as well. Findings in this study demonstrated that classroom observation scores is a more eliminative criterion in terms of discrimination. Hence, classroom observation scores can be addressed as additional criterion by researchers who seek to develop higher quality measurement instruments. Therefore, it may be possible to obtain information about items that are related to both teachers' classroom observations and the total test score. Accepting test scores as the sole criterion will raise doubts since the criterion is restricted and momentary. Baki and Birgin (2002) investigated portfolios as alternative assessment tools in Mathematics education and reported that traditional assessment and evaluation tools provided restricted information about students with no depth and that more dynamic assessment and evaluation methods were needed to present students in more detail.

Since item discrimination index is a convergent validity index, it is vital to prove this validity using scientific methods. One of these methods may include teachers to score classroom observations to check concordance with this criterion. In their study that investigated the scale development and adaptation studies published in Psychology and Educational Sciences Journals in Turkey, Çüm and Koç (2013) reported that very few scale development studies included empirical validity studies.

Since this study was conducted in the field of Mathematics, it can be suggested to replicate it in other lessons and subject areas. Concordance between item and test scores is expected to decrease when the scope is broadened, therefore, the study may be replicated with more restricted and more comprehensive tests to compare the obtained results.

References

Airasian, P. W. (1979). A perspective on the uses and misues of standardized achievement tests. *NCME Measurement in Education, 10*(3), n3.

Anderson, P. S. (1987). Comparison of student attitudes about seven formats of educational testing, with emphasis on the MDT multi-digit testing technique.Paper presented at the Annual Meeting of the Mid-Western Educational Research Association, Chicago, II. (ERIC Document ED 295999).

Arıcı, H. (2004). *Statistics: Methods and Application*. Hacettepe University Press.

Baki, A., & Birgin, O. (2002). İndividual Development File Application as an Alternative Assessment in Mathematics Education. *V. National Science and Mathematics Education Congress Book*, II, (913-920). Ankara.

Baykul, Y. (2000). *Measurement in Education and Psychology: Classical Test Theory and its Application.* Ankara: ÖSYM, Turkey.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

Çüm, S., & Koç, N. (2013). Investigation of scale development published in the journal of psychology and educational sciences and adaptation in Turkey. *Journal of Educational Sciences & Practices, 12*(24).

DeMars, C. (2010). *Item Response Theory*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195377033.001.0001

Demir, İ. (2011). Identity function scale: Turkish validity and reliability. *Educational Sciences in Theory and Practic*e, *11*(2), 571-586.

Gipps, C. V. (1994). *Quality Assurance in Teachers' Assessment*. In W Harlen (Ed) *Enhancing Quality in Assesment.* London. Paul Chapman

Herman, J. L., & Dorr-Bremme, D. W. (1983). Uses of testing in the schools: A national profile. *New Directions for Testing & Measurement, 19*, 7-17.

Kagitçibaşi, Ç. (1976). *Human and People: Introduction to Social Psychology.* Ankara: Turkish Social Sciences Association.

Kilmen, S. (2014). *Basic Concepts in Measurement and Evaluation*. N. Çıkrıkçı-Demirtaşlı (Ed.). *Measurement and Evaluation in Education* (s.34-68). Edge Academy. Ankara.

Kızılkaya, G., & Aşkar, P. (2009). Development of a reflective thinking skill scale for problem solving. *Education and*

Science, *34*(154), 82-92.

Parkes, C., & Maughan, S. (2009). Policy and Research Seminar on Methods for Ensuring Reliability of Teacher Assessments. Proceedings of National Foundation for Educational Research and Chartered Institute of Educational Assessors (The Royal Institute of British Architects, London, UK, June 2, 2009). *National Foundation for Educational Research*.

Salmon-Cox, L. (1981). Teachers and standardized achievement tests: What's really happening?. *The Phi Delta Kappan*, *62*(9), 631-634.

Stiggins, R. J., & Bridgeford, N. J. (1985). Performance assessment for teacher development. *Educational Evaluation and Policy Analysis*, *7*(1), 85-97. https://doi.org/10.3102/01623737007001085

Tekin, H. (1996). *Measurement and Evaluation in Education.* Yargi Publications, Ankara.

Turgut, M. F. (1983). *Program Evaluation: Education in republican period*, Milli Eğitim Publication. İstanbul.