International Journal of Evaluation and Research in Education (IJERE)
Vol.7, No.1, March 2018, pp. 32~38
ISSN: 2252-8822, DOI: 10.11591/ijere.v7.i1.11488

Effects of Test Level Discrimination and Difficulty on Answer-Copying Indices

Onder Sunbul, Seha Yormaz
Department of Measurement and Evaluation in Education, Mersin University, Turkey

Article Info

Article history:

Received Nov 19, 2017
Revised Jan 27, 2018
Accepted Feb 14, 2018

Keyword:

Answer copying
Cheating
Copy detection
GBT
Omega (ω)

ABSTRACT

In this study Type I Error and the power rates of omega (ω) and GBT (generalized binomial test) indices were investigated for several nominal alpha levels and for 40 and 80-item test lengths with 10,000-examinee sample size under several test level restrictions. As a result, Type I error rates of both indices were found to be below the acceptable nominal alpha levels. The power study showed that average test difficulty was very effective for power (true detection) rates of indices. Clear patterns were observed for the increase of test difficulty in favor of both ω and GBT power rate. Contrary to expectations; average test discrimination was not as effective as average test difficulty. The results of the interaction effects of item discrimination and difficulty showed that for the cases whose b parameters were lower than 0 with weak discrimination, indices had weak power for both ω and GBT. In addition, for the cases whose b parameter levels were below zero with high discrimination indices, the power performance of both answer-copying indices were very weak. Results for test length showed that with the increase of test length the power rate of both ω and GBT tended to increase. Also, ω performed slightly better than GBT or very close to GBT for 80-item test length however, ω performed better than GBT in terms of power rate for the cases with 40-item test length.

Corresponding Author:

Onder Sunbul,
Faculty of Education, Department of Measurement and Evaluation in Education,
Mersin University,
Mersin, Turkey.
Email: ondersunbul@mersin.edu.tr

1. INTRODUCTION

Multiple choice item tests have been used for decades for decision making in high-stake tests. Despite the fact that multiple choice item tests have numerous advantages for large scale examinations; they also have several weaknesses which especially threaten test security issues. Test security is vital for valid decisions that depend on test scores. With the increase of population in countries, competition among examinees or educational organizations to cover a requirement, to pass a cutoff point, to find a job has increased. The increase of population across limited sources forces individuals or educational organizations to cheat in examinations. Cheating in examinations can be conducted in several ways: Transmitting answers of questions, memorizing questions for future examinees which result in item exposure, collusion during examinations, erasures after examinations, obtaining questions before examinations to have pre-knowledge about items etc. Technological developments have made communication easier. However, one of the side effects of these improvements is the threats for test security. When we compare it with the last decades nowadays it is very easy to transmit or broadcast the answers of an examination. Test security area, which is a subdomain of educational measurement, has to improve itself because of these threats. Literature review presents numerous cheating detection procedures [1]-[23]. As previously mentioned, there are several ways to cheat in tests; however, this study only focuses on answer copying. Several answer-copying detection

indices focus on matching examinees' incorrect responses and some of them focus on matching both examinees' correct and incorrect responses. The performances of answer-copying detection statistics may vary depending on assumptions of indices as well as the test and item specifications. In other words, decisions made for examinees for answer copying must be evaluated with paying attention to assumption of indices, test and examinee specifications as an additional evidence of cheating. Literature review shows that Omega (ω) and Generalized Binomial Test (GBT) indices are quite remarkable.

Omega (ω) index was proposed by Wollack [23]. Briefly, ω compares the expected match with the observed match between two examinees' response vectors. Original formula of ω is based on item parameters obtained from the nominal response model [24] of Item Response Theory. The expected values of each response alternative will be estimated by using nominal response model estimations under the restriction of ability parameters of probable copier and source. Despite the fact that ω's original formula was developed by using Nominal Response Model and using option matrix of a multiple choice item test, it is also possible to compute ω on dichotomous data sets by using 1PL, 2PL or 3PL models of IRT [25],[26]. Literature review presents several studies which compare ω for nominal responses and dichotomous responses. Because of the hardiness of using or working on option matrix, researchers in security companies tend to work on dichotomous responses rather than nominal responses [27].

Generalized Binomial Test is another index developed by van der Linden and Sotaridona [28] and used for answer-copying detection. Like - ω index, it takes into account identical correct and incorrect answers of a suspected pair of copier and source examinee. GBT uses compound binomial distribution. This test also requires estimations obtained from IRT models (Nominal Response Model or 1PL, 2PL or 3PL models of IRT). In addition to distributional differences between GBT and ω, another distinction between them is pointed by Zopluoglu [29] that GBT is a similarity statistics (symmetric) however ω is an answer-copying index (asymmetric). There are very limited studies in literature for GBT.

1.1. Purpose of the Study

Numerous studies have been conducted to compare answer-copying detection indices for several conditions in terms of Type I error and power studies [5]-[7],[12],[13],[15],[19],[20],[23],[25],[26],[28],[30]-[33]. Literature review points that two of the answer copying indices, ω [23] and GBT [28], are remarkable with low Type I error rates and high detection power ratios [19],[20],[23],[25],[26],[30]-[34]. However there are limited studies which compare the performance of these indices in literature. Zopluoglu and Davenport [26] investigated the performance of ω and GBT for several amount of copying and copier & source ability crossing. Zopluoglu [25] investigated the performance of ω, GBT, K, $\overline{K}_1$, $\overline{K}_2$, S_1, S_2 on several simulated data sets and real data set. Sunbul and Yormaz [34] investigated the performance of ω index on the 80-item five option multiple choice test with 10,000 examinees. The data set contains four different item types: 20 easy - low discriminative items, 20 hard - low discriminative items, 20 easy - high discriminative items and 20 hard -high discriminative items. In their study they investigated the performance of ω for cheating over specified item groups. Literature review presents very limited studies which compare the performances of ω and GBT by taking into account test and item properties and previous studies show that item properties and especially difficulty levels affect the performance of indices [25].

Literature review shows no study for performance investigation on test level where tests have homogeneous item groups in terms of item difficulty and discrimination. In this study, we generated 12 different homogeneous (in terms of item properties) item groups for 40 and 80-item test lengths. By crossing homogeneous item groups with test lengths; 24 different dichotomous (1-0) test data were obtained and we observed the performance of ω and GBT on several test data sets. Main purpose of this study was to investigate performance of ω and GBT for answer-copying detection on different tests in terms of average item difficulty and discrimination.

2. RESEARCH METHOD

2.1. Research Design

Dichotomous (1-0) data were generated under the framework of 2PL model of Item Response Theory by using R programming language. To manipulate test average item difficulty; four different b parameter ranges (-2.00 – -1.10, -1.00 – -0.10, 0.00 – 0.90 and 1.00 – 2.00) were used. To manipulate test average item discrimination; three different a parameter ranges were used (0.40 – 0.79, 0.80 – 1.19 and 1.20 – 1.60). a and b parameters were sampled from uniform distributions depending on the related minimum and maximum values of relevant parameter ranges. Based on a and b parameter restrictions, dichotomous item responses were generated for 40-item and 80-item test length for 10,000-examinee sample size.

Random item copying procedure was applied for fixed copying amount of 40% for each replication. Amount of copying was taken and fixed as 40% for each condition because previous studies [23],[26][30],[31] showed that power rates of indices were quite satisfactory for 30% and more.

Ability levels of copier and source are important for the performance of answer-copying indices. Several studies in the literature have pointed to the importance of examinee ability levels on answer copying [26],[34]. For this study the abilities of examinees were obtained by using irtoys [35] R package depending on 2PL Item Response Model. For each replication, copiers were sampled from first quartile of ability distribution and sources were sampled from fourth quartile of ability distribution. It is assumed that low ability level examinees tend to cheat from high ability level examinees.

Four different nominal alpha levels (0.001, 0.005, 0.01 and 0.05) were used for Type I error and power study. Five hundreds replications were conducted for each experimental cell of condition crossing. Omega and GBT indices were evaluated in terms of Type I error and power study. Table 1 presents simulation design regarding research conditions and its levels.

Table 1. Simulation Design Conditions and Levels

Condition	Number of Levels	Level Values
Sample Size	1	10000
Test Length (*T*)	2	40 (*T1*) 80 (*T2*)
Amount of Copying	1	40%
Item Discrimination (*a*)	3	0.40 – 0.79 (*a1*) 0.80 – 1.19 (*a2*) 1.20 – 1.60 (*a3*)
Item Difficulty (*b*)	4	-2 – -1.1 (*b1*) -1.0 – -0.1 (*b2*) 0.0 – 0.9 (*b3*) 1.0 – 2.0 (*b4*)
Indices	2	ω and GBT
Number of Replications	500	

In the current study, conditions and its levels were determined by paying attention to the conditions of researches [19]-[20],[23],[26],[30] to make comparisons. The empirical Type I error rate was computed by determining the ratio of honest examinee pairs which were flagged as copier incorrectly by indices over 500 replication (honest pairs). For the power study, power of indices was computed by determining the ratio of examinee pairs who were flagged as copier correctly by indices over 500 replication. In contrast with Type I error study, copying scenarios were implemented to the examinee pairs' (copier -source) response vectors. To create copied examinee pairs, the fixed amount 40% of the random item responses of copier examinee was altered with the responses of source examinee for each experimental cell and replication. CopyDetect package of R [36] was used for computing the probabilities of indices. Both for Type I error and power study results were aggregated by specified nominal alpha levels.

3. RESULTS AND ANALYSIS

3.1. Type I error study

Type I error rates were given in Table 2. The results of the Type I error study showed that, most of the Type I error rates were zero or very close to zero. Any Type I error inflation for any condition couldn't be observed regardless of test average difficulty and discrimination. Most of the results were 0.00 and several discrepancies from 0.00 were observed for ω index. For the most difficult test (b4), only one exceedance was observed for nominal alpha level 0.05 for ω index. GBT performed very well for each condition in terms of Type I error rate.

Table 2. Type I Error rates of ω and GBT for Simulation Conditions

	Alpha Level	T1 b1 ω	T1 b1 GBT	T1 b2 ω	T1 b2 GBT	T1 b3 ω	T1 b3 GBT	T1 b4 ω	T1 b4 GBT	T2 b1 ω	T2 b1 GBT	T2 b2 ω	T2 b2 GBT	T2 b3 ω	T2 b3 GBT	T2 b4 ω	T2 b4 GBT
a1	0.001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.01	0	0	0	0	0	0	0	0	0	0	0.01	0	0	0	0.01	0
	0.05	0.01	0.01	0.03	0	0.04	0.01	0.04	0	0.01	0.01	0.04	0.02	0.03	0.01	0.06	0.01
a2	0.001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0
	0.05	0	0	0.02	0.01	0.04	0.01	0.03	0	0	0	0.01	0.01	0.03	0	0.04	0
a3	0.001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.005	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0.01	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.01	0
	0.05	0	0	0	0	0.01	0	0.02	0	0	0	0	0	0.01	0	0.03	0

3.2. Power study

It is observed in Figure 1 that the power of ω and GBT indices tended to increase with the increase of test length and this result is consistent with the literature [19],[20],[23],[30]-[32]. Another clear pattern was observed for increases in test difficulty. With the increase of b parameters of test items, power of ω and GBT indices tended to increase, too. This result is also consistent with previous studies in terms of the relationship between the difficulty and copy detection power [25],[34]. However a clear increase pattern couldn't be observed with the increase of a parameters of test items. The difference between ω and GBT indices increased in favor of ω with the increase of a parameters especially for the tests with harder items. Another inference may include the interaction between test length and item difficulty. Both ω and GBT indices were observed to be more capable to detect copiers when they cheated in longer tests with harder items especially for 0.05 and 0.01 nominal alpha levels.

When we compare the results of ω and GBT, ω was found to perform better than GBT nearly for all conditions. ω index clearly performed better than GBT especially for the 40-item test length whose items' b parameters were higher than 0.00. The difference between the power rates of two indices got closer with the increase of test length. In general, the performance of ω and GBT indices were unsatisfactory for tests whose items' b parameters were lower than 0.00 for all nominal alpha levels

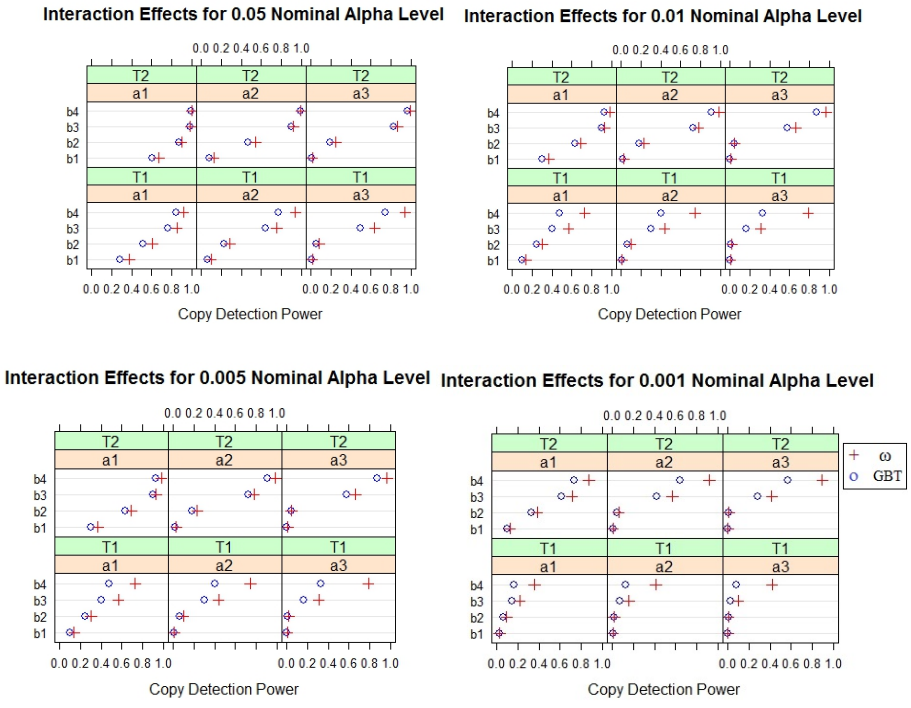

Figure 1. The Interaction Effects of Conditions for Power Study

4. DISCUSSION

The studies [19],[20],[23],[25],[26],[30]-[34] showed ω index to have lower Type I error rate and high detection power for answer copying under many conditions. Literature review presents that answer-copying detection power of ω tended to increase with the increase of test length. The results of this study were found to be similar to those in literature in terms of low Type I error rates and power increase with the increase of test length.

Zopluoglu and Davenport [26] compared the performances of ω and GBT with crossing the ability levels of copier and source and showed that GBT was slightly better than ω in general. In contrast, present study showed that ω performed better than GBT in terms of answer copy detection power especially for 40-item test length with harder items.

Sunbul and Yormaz [34] reported several Type I error inflations for low ability level source for ω index. They also reported satisfactory power rates except when low ability level copiers cheated from high discriminative and easy items. The results of the current study were found to be consistent with the results of Sunbul and Yormaz's study [34] as well. Cheating in easy items might be undetectable by answer-copying detection indices. Another important point to be cited is the fact that the current study tried to manipulate item difficulty and discrimination in test level. In other words, homogeneous tests were generated in terms of difficulty and discrimination. However Sunbul and Yormaz [34] generated an 80 item heterogeneous test which covered several item types (easy-hard, more discriminative-less discriminative) in a single test since they were interested in cheating from specific types of items. The main reason for this might be the amount of item information which was provided by item information function for the easy items.

5. CONCLUSION

Both of ω and GBT performed well nearly for all conditions in terms of Type I error rates. The results of power study for the both of ω and GBT showed that they performed well for tests which have harder items however the results for the tests which have easy items were unsatisfactory. Detecting cheating from easy item tests might be a problem. In terms of real life applications, it is not common to cheat from easy items since it might not be a big gap for answer detection indices. However practitioners and researchers should be more careful while investigating answer copying on easy tests. More studies are needed to invesitgate the effect of item information amount for answer copying.

ACKNOWLEDGEMENTS

This study is the revised and improved version of the oral presentation in V. Congress on Measurement and Evaluation in Education and Psychology 2016 as Investigating the Performance of Copy Detection Indices According to Item Discrimination and Difficulty

REFERENCES

[1] Angoff W. H., "The development of statistical indices for detecting cheaters," *Journal of American Statistical Association,* vol. 69, pp. 44-49, 1974.
[2] Anikeef A. M., "Index of collaboration for test administrators," *Journal of Applied Psychology*, vol. 38, pp. 174-177, 1954.
[3] Argenal R. N., *et al.*, "A new index for detecting collusion and its statistical properties," National Convention on Statistics, 2004.
[4] Armstrong R. D. and Shi M. A., "Parametric cumulative sum statistic for person fit," *Applied Psychological Measurement*, vol/issue: 33(5), pp. 391-410, 2009.
[5] Bay M. L. G., "Detection of copying on multiple-choice examinations," [Doctoral dissertation], Southern Illinois University, 1987. *Dissertation Abstracts International*, vol/issue: 56(3-A), pp. 899, 1994.
[6] Bellezza F. S. and Bellezza S. F., "Detection of cheating on multiple-choice tests by using error-similarity analysis," *Teaching of Psychology,* vol. 16, pp. 151-155, 1989.
[7] Belov D. I., "Detection of answer copying based on the structure of a high-stakes test," *Applied Psychological Measurement*, vol/issue: 35(7), pp. 495–517, 2011.
[8] Belov D. I. and Armstrong R. D., "Automatic detection of answer copying via kullback-leibler divergence and K-index," *Applied Psychological Measurement,* vol/issue: 34(6), pp. 379–392, 2010.
[9] Bird C., "The detection of cheating in objective examinations," *School and society,* vol. 25, pp. 261-262, 1927.
[10] Bird C., "An improved method of detection cheating in objective examinations," *Journal of Educational Research,* vol. 25, pp. 261-262, 1929.
[11] Clark J. M., "Aberrant response patterns as a multidimensional phenomenon: Using factor-analytic model comparison to detect cheating," ProQuest LLC, University of Kansas, 2010.

[12] Frary R. B., *et al.*, "Indices of cheating on multiple-choice tests," *Journal of Educational Statistics,* vol. 6, pp. 152-165, 1977.
[13] Hanson B. A., *et al.*, "A comparison of several statistical methods for examining allegations of copying," Iowa City, IA: American College Testing, 1987. ACT Research Report Series No. 87-15.
[14] Harpp D. N., *et al.*, "Crime in the classroom – Part II, an update," *Journal of Chemical Education,* vol/issue: 73(4), pp. 349-351, 1996.
[15] Holland P. W., "Assessing unusual agreement between the incorrect answers of two examinees using the K-index: statistical theory and empirical support," Princeton, NJ, Educational Testing Service, 1996. Research Report No: 94-4.
[16] Hui H. F., "Stability and sensitivity of a model-based person-fit index in detecting item pre-knowledge in computerized adaptive test," Dissertation Abstracts International Section A: Humanities and Social Science, University of Hong Kong, 2010.
[17] Saupe J. L., "An empirical model for the corroboration of suspected cheating on multiple-choice tests," *Educational and Psychological Measurement,* vol. 20, pp. 475-489, 1960.
[18] Shu Z., "Detecting test cheating using a deterministic, gated item response theory model," [Doctoral dissertation], The University of North Carolina at Greensboro, 2010. Dissertation Abstracts International Section A: Humanities and Social Sciences, 2011.
[19] Sotaridona L. S. and Meijer R. R., "Statistical properties of the K-index for detecting answer copying," *Journal of Educational Measurement,* vol. 39, pp. 115-132, 2002.
[20] Sotaridona L. S. and Meijer R. R., "Two new statistics to detect answer copying," *Journal of Educational Measurement,* vol. 40, pp. 53-69, 2003.
[21] Sotaridona L. S., *et al.*, "Detecting answer copying using the kappa statistic," *Applied Psychological Measurement,* vol. 30, pp. 412-431, 2006.
[22] van der Linden W. J. and Sotaridona L. S., "A statistical test for detecting answer copying on multiple-choice tests," *Journal of Educational Measurement,* vol. 41, pp. 361-378, 2004.
[23] Wollack J. A., "A nominal response model approach to detect answer copying," *Applied Psychological Measurement,* vol. 21, pp. 307-320, 1997.
[24] Bock R. D., "Estimating item parameters and latent ability when responses are scored in two or more nominal categories," *Psychometrika,* vol. 46, pp. 443-459, 1972.
[25] Zopluoglu C., "Classification performance of answer-copying indices under different types of irt models," *Applied Psychological Measurement,* vol/issue: 40(8), pp. 592-607, 2016.
[26] Zopluoglu C. and Davenport E. C., "The empirical power and type I error rates of the GBT and ω indices in detecting answer copying on multiple-choice tests," *Educational and Psychological Measurement,* vol/issue: 72(6), pp. 975-1000, 2012.
[27] Assessment Systems, "User's manual for SIFT: Software for investigating fraud in testing," 2016.
[28] van der Linden W. J. and Sotaridona L. S., "Detecting answer copying when the regular response process follows a known response model," *Journal of Educational and Behavioral Statistics*, vol. 31, pp. 283-304, 2006.
[29] Zopluoglu C., "Similarity, answer copying, and aberrance: Understanding the status quo," in G. J. Cizek and J. A. Wollack, "Handbook of quantitative methods for detecting cheating on tests," New York, NY: Routledge, pp. 25-46, 2017.
[30] Wollack J. A., "Comparison of answer copying indices with real data," *Journal of Educational Measurement,* vol. 40, pp. 189-205, 2003.
[31] Wollack J. A., "Simultaneous use of multiple answer copying indexes to improve detection rates," *Applied Measurement in Education,* vol. 19, pp. 265-288, 2006.
[32] Wollack J. A. and Cohen A. S., "Detection of answer copying with unknown item and trait parameters," *Applied Psychological Measurement,* vol. 22, pp. 144-152, 1998.
[33] Wollack J. A., *et al.*, "Defining error rates and power for detection of answer copying," *Applied Psychological Measurement,* vol. 25, pp. 385–404, 2001.
[34] Sunbul O. and Yormaz S., "Investigating the performance of ω index according to item parameters and ability levels," *Eurasian Journal of Educational Research,* in press, 2017.
[35] Partchev I. I., "Simple interface to the estimation and plotting of IRT models," R package version 0.1.7, 2014.
[36] Zopluoglu C., "CopyDetect: Computing statistical indices to detect answer copying on multiple-choice tests," R package version 1.2, 2016.

BIOGRAPHIES OF AUTHORS

Onder Sunbul is an Assistant Professor in Mersin Universty in the Department of Measurement and Evaluation in Education. He recived his MSc and PhD degree from Department of Measurement and Evaluation in Education, Mersin University. His main interests are Educational Measurement, Computational Psychometrics, Test Security, Item Response Theory, and Statistical Programming.

SehaYormaz is a PhD student. She was born in Mersin on March 6, 1983. She obtained her BCs degree (2005) in Ankara at Hacettepe University, Faculty of Education, and MSc degree (2014) at Mersin University, Department of Measurement and Evaluation in Education. Currently she is a PhD student at Mersin University, Department of Measurement and Evaluation in Education and works as a science teacher in Mersin, Turkey. Her main interests are CTT, IRT, R programming and cheating.