ISSN: 2790-4407

Journal of Computer Adaptive Testing in Africa

2023; 2; 1-9

Examining item compromise: an introduction to deterministic gated item response theory model (DGIRT)

Yusuf Olayinka Shogbesan

Department of Arts and Social Sciences Education, Al-Hikmah University, Ilorin, Nigeria

ABSTRACT

The calibration of item parameters (difficulty, discrimination and guessing parameters) estimate may only consider the true and not the cheating abilities of examinees. In an effort to detect the occurrence of test cheating due to compromise in multiple-Choice items, the Deterministic Gated Item Response Theory Model was developed to provide information about cheating effectiveness of examinees, measure the extent of item fit for the compromised items, assess the sensitivity and ascertain the specificity to detect cheating due to item compromise. Hence, the model was meant to provide information on the extent of which the item response theory psychometric estimates are sensitive to item compromise when cheating occurs in large-scale examinations. This paper examine the concepts of cheating, item compromise and provides a brief overview of the Deterministic Gated Item Response Theory Model (DGIRT). It was recommended among others that Psychometricians should consider the validation of Deterministic Gated IRT model and other new IRT models that will account for the cheating ability of examinee unlike the "normal" IRT model that produces the probability of an item response for varying values of θ (ability).

Keywords: tests compromise, score inflation, cheating, deterministic gated item response theory model (DGIRT)

* Corresponding author.

Department of Arts and Social Sciences Education, Al-Hikmah University, Ilorin, Nigeria yoshogbesan@alhikmah.edu.ng

To cite this article:

Shogbesan, Y. O. (2023). Examining item compromise: an introduction to deterministic gated item response theory model (DGIRT). *Journal of Computer Adaptive Testing in Africa, 2*, 1-9.

Received: 2 November 2023
Accepted: 30 November 2023
Published: 19 December 2023

INTRODUCTION

Tests are meant to elicit information about a latent ability of an individual and to provide evidence so that educational decisions can be made about the individual. These decisions, when made, provide information about students on whether they have reached a particular level of skill and knowledge. It may be used to evaluate a teaching programme or to make decisions about the next aspect of teaching for particular students. Although several schools of thought have argued for and against the use of tests, it has been the best available and mostly used instrument of measurement (Afolabi, 2012). The arguments against the use of tests may be as a result of various factors that affect test score reliability, validity and usability. This effect as it relates to the test itself can be as a result of different standard, motivational factors, familiarity with test items, bias, cheating as a result of tests items being compromised as well as other related examination malpractices.

Item compromise is a form of test security breaches which may undermine the psychometric properties of tests. It can be defined as occurring when evidence exists that an item performance has changed during some defined time span and it is reasonable to believe that the performance changes are due to its content having been distributed beyond defined valid usage boundaries or due to over exposure to test takers (Zara & Pearson, 2006). Hence, an item can be regarded to be compromised specifically when test-takers have access to the test questions prior to the time the test is scheduled to hold or before completing the test (Drasgow et al., 2009; Lievens & Burke, 2011). Item compromise is regarded as a form of cheating in examinations because it is an act, which gives a student undue advantage over other students in answering any set of compromised examination questions.

Stakeholders of various tests commonly recognize test cheating, as a negative factor invalidating the inference made based on tests. Testing agencies or other stakeholders have suffered great pain from diffusedly existing test cheating, especially in high-stake tests (Shu, 2010). Cheating in examinations have become a serious threat to the reliability and validity of examinations. It is regarded as any action that violates the rules for administering a test, any behaviour that gives an examinee an unfair advantage over other examinees, or any action on the part of an examinee or test administrator that decreases the accuracy of the intended inferences arising from the examinee's test score or performance (Cizek, 2001).

The incidence of cheating can be caused by breach in security of examination questions which have been seen as a major problem threatening public examination bodies in Nigeria and have impacted negatively on the integrity of public examinations over the years (Ojerinde, 2015). This breach in security of examination questions can be as a result of leakages in the questions prior to the commencement of the examination or due to deliberate item exposure to the tests-takers before the commencement of the examination. Furthermore, the occurrence of examination insecurity can also manifest in various forms such as leakages of questions, impersonation, swapping examination document, spying/coping from prepared answers, use of unauthorized calculator or similar electronic devices, smuggling out of question papers/answer sheets among others. In addition to occurrence of cheating that have always been witnessed in testing contexts where there are substantial consequences for individuals, the phenomenon of organised and professional cheating has arisen more recently as one of the unintended consequences of credentialing programmes in which examination requirements often play a prominent role in licensure or certification decisions (Wollack & Cizek, 2017). Moreover, test-takers having prior knowledge of specific test content are an indication that the test items have been compromised either as a result of frequent use of test item in testing (repetition), item cloning, item rotation, item overuse, or test item leakages or breaches specifically to a situation in which test-takers have access to test questions prior to completing the test (Drasgow, Nye, Guo, & Tay, 2009; Lievens & Burke, 2011).

In the Nigeria educational sector and assessment situations, the phenomenon of organized and professional cheating have often occurred at examination centres usually tagged "special centres". However, Computer-Based Testing have been introduced by some examination bodies in Nigeria (e.g. JAMB) and used as a measure to combat examination insecurity as most of the cases of examination malpractice are reported to be mostly carried out with the paper - pencil testing mode (Ojerinde, 2015; Idika, 2015). According to the Registrar of the Joint Admissions and Matriculation Board, no fewer than 178 cases of malpractices were detected in the Unified Tertiary Matriculation Examination during the 2022 and 2023 batches of the examination with a total of 94 cases were detected during the 2022 UTME, while 84 cases were detected in 2023 (https://punchng.com/jamb-detected-178-malpractices-in-2022-23-report/, 4th July 2023). In recent time, JAMB in the 2023 national examination was able to curb even cases of post examination malpractice. The detected was as a result of the integration of technology to administer tests and for credentialing. Although the application of cutting edge technology has been of immense help; ironically, the same technology has proved a

challenge in that it has also been available for the examination cheats to deploy (https://tribuneonlineng.com/jamb-mmesoma-and-the-new-face-of-examination-malpractice/,9th July, 2023).

Cheating, occurring at the pre-testing stage, can be as a result of test item being compromised or exposed to tests takers. This situation will allow for test-takers to harvest compromised items during testing. Specifically, when test takers are familiar with test items as a result of the item being compromised due to item pre-knowledge, it may affect positively their scores on such test especially if properly haversted. According to Shu, Henson and Luecht (2013), most organisations involved in high-stakes testing are fairly certain that test cheating or compromise (usually by item over-exposure and memorization) widely occurs, especially when many items and test forms must be exposed to accommodate on-demand computer-based testing (CBT). The occurrence of item exposure stem from the seating capacity, scheduling limitations and item development costs (Luecht, 1998, 2005; Drasgow, Luecht, & Bennett, 2006; Shu, Henson & Luecht, 2013).).

Hence, when cheating occurs due to item compromise, estimates of an examinee's performance are no longer accurate, perhaps the most obvious example of cheating as a threat to validity occurs when an individual is unduly advantaged and receives a score that is higher than his or her true estimate (Klapetzky, 2012). Moreover, for tests to maintain the characteristics of being fair, reliable, and valid, the tests items need to be secured because decisions based on scores affected by cheating are unacceptable.

However, in a situation where there is a breach in examination security leading to a leakage in the examination questions, there is an undue access to the tests item pool leading to a compromise as test takers may memorize the test items and subsequently reproduce them during testing situations. As a result, individuals with prior knowledge of some items may use that information to inflate their test scores (McLeod & Schnipke, 1999). This situation may also lead to the occurrence of test bias. Students' scores have been found to be inflated on compromised and practised test items which had contributed to the score invalidity. The impact of compromised anchor items on IRT equating under the nonequivalent anchor test design have been found to artificially increase the equated scores of the entire examinee group that was administered the compromised form (Jurich, DeMars, & Goodman, 2012; Idika, Shogbesan & Ogunsakin, 2016). Also, item compromise have occurred when it is evidence

that an item performance has changed during some defined time span due to its content having been distributed beyond defined valid usage boundaries or due to over exposure to test takers (Zara & Pearson, 2006). The inflated score would essentially be a misrepresentation of that individual's performance, thus yielding an inaccurate estimate of performance. Then, it may be true that item compromise often presents a more tractable set of score invalidity issues especially as its affects the parameter estimates.

According to Zimmermann, Klusmann and Hampe (2016), if an item is compromised due to pre-knowledge, changes are to be expected in the item parameter estimates which are generated by an Item Response Theory (IRT) model. The item difficulty parameter should decrease because pre-knowledge produces an excess of correct responses. Similarly, item discrimination should decrease because, in the sub-group of pre-knowing, test takers already almost have all correct items pre-known in advance. Therefore, the items are not related to ability as estimated by the uncompromised part of the item set. Although various statistical cheating detection models have been developed to provide information on the occurrence of test cheating (e.g., similarity indices, answer copying indices, person fit indices) which rely on aggregation of individual statistics (Segall, 2002), the Deterministic, Gated IRT (DGIRT) model derives test cheating or test compromise summaries by response matrix (Shu, 2010).

Deterministic, Gated IRT (DGIRT) model: An Overview

The traditional IRT models can be used to describe the relationship between items and ability for either dichotomous models; used for items that place responses in two categories e.g. correct/incorrect or polytomous models used for items that place responses in more than two categories e.g. items scored for partial credit (Cohen et al., 2001). Within the general IRT framework, many model shave been formulated and applied to real test data. For dichotomous items, the 1, 2, and 3 parameter logistic models are most common (1PL, 2PL, 3PL), and models including an upper asymptote parameter (e.g., 4PL) are also possible. However, for polytomous items, variations of the Partial Credit Model, Rating Scale Model, Generalized Partial Credit Model, and Graded Response Model are available for ordered responses, and the Nominal Model is appropriate for items with a non-specified response order. All these models are the cornerstone of IRT; they are the pivots upon which the theory depends and they reveal information about the latent behavior of the items and the examinee which make it easy for measurement community to make right predictions (Ogunsakin & Shogbesan, 2018).

However, McLeod, Lewis and Thissen, (2003) opined that given the "normal" IRT model produces the probability of an item response for varying values of θ (ability) while an "item pre-knowledge" will modify the IRT model because it will now be that the probability of a correct response to an item is the combination of the probability of answering the item correctly based on the test taker's pre-knowledge of the item and the probability of answering the item correctly based on the test taker's underlying proficiency in the case that the test taker did not have pre-knowledge of the specific item. Hence, the Deterministic, Gated Item Response Theory Model (DGIRT; Shu, 2010) was proposed to detect test cheating that results from item over-exposure. Specifically, this model addresses cheating that has occurred because the examinees have had previous access to an item. This model seems to be an attractive and promising tool for practitioners to respond to test cheating by both individual and organized cheating, given its modeling design, the provided information, sensitivity and specificity level (Shu, 2010).

The DGIRT classifies test takers as cheaters or non-cheaters by conditioning on two mutually exclusive item types. The first type of item is one that has probably been compromised. This first type of item could be identified based on empirical exposure counts, time in use, or other indicators (called "exposed items"). The second type of item is considered a secure item due to its recent release or other factors (called "unexposed items") (Segall, 2002; Shu, Henson & Luecht, 2013).

The DGIRT model provides information about the cheating probability for each single examinee. It also characterises cheaters' real knowledge level by (θt) and the severity of their cheating activities (θc), which enable practitioners to have deeper inside analysis on the characteristics of cheaters and non-cheaters. Also, when cheating occurs, the ability of the examinee will not reflect true ability as the probability of correct response to an item now depend on both the true ability and cheating ability of the examinee. The cheating ability of the examinee is characterised by the examinees ability to properly use the undue advantage to ensure a correct response to an item. The cheating effectiveness can be affected by the nature of the cheating condition and cheating size. These situations are expected to affect the item fit and item parameter estimates. Furthermore, the item fit to the IRT model should decrease because compromised items will be frequently correct when the 2PL model predicts an incorrect response (Shu, 2010, 2013; Zimmermann, Klusmann & Hampe, 2016). From the above explanation, it implies that the item parameters of the test are affected when a compromise situation

occurs. As such, it is therefore imperative to determine the sensitivity of the DGIRT model to such compromise condition.

However, this model is designed to mainly detect test cheating by item-preview, item memorization or internet-collaboration. It is a Rash-Model based which does not take guessing and item discrimination into account and proves to be useful in large-scale tests where a large number of examinees and items are available (Shu, 2010). The Deterministic, Gated IRT model is not only a cheating model which can only be applied in detecting test cheating. It can be seen as a general model to distinguish two groups of examinees and be able to characterise the two groups of examinees with two latent traits.

However, several researchers have develop statistical approaches to detect the occurrence of cheating. This include among others, a Bayesian method to detect cheating due to item pre-knowledge, the investigation of changes in item difficulty, item discrimination, item fit and local dependence between non-compromised and compromised data sets. Despite the development of various statistical or IRT approaches to identify cheating on compromised items which can be caused by item pre-knowledge among others, empirical evidence showed that even highly sophisticated statistical methods are not really effective to detect cheating especially if the proportion of potentially compromised items is large relative to the item set guaranteed to be uncompromised and also if the number of affected items and the number of cheating respondents is small (Shogbesan, 2021).

Although the DGIRTM successfully overcomes some limitations of the previous cheating detection techniques, this new model is certainly not immune to abuse or probative misuse. As such, the successful and appropriate application in different real settings is the ultimate purpose. Hence, practitioners are therefore urged to evaluate the effectiveness of the Deterministic, Gated IRT model.

CONCLUSION AND RECOMMENDATION

1. Psychometricians should consider the validation of Deterministic Gated IRT model and other new IRT models that will account for the cheating ability of examinee unlike the "normal" IRT model that produces the probability of an item response for varying values of θ (ability).
2. The Deterministic, Gated IRT model as a statistical method to detect cheating was developed and used within a stimulation design and they are to be validated empirically using real data.

3. Computerized Adaptive Testing should be employed as it will avoid the use of repeated items directly for examinees as only items related to their ability level will be calibrated for them during examination. It also help to ensure item exposure control since every examinee will receive an item different from the one another depending on their ability levels.
4. Test item security should be seriously considered as a critical part of the factors that affects validity of tests scores as such subsumed under the validity issues needed to be explored statistically to provide evidence that scores obtained passed the integrity test.

REFERENCES

Afolabi, E. R. I. (2012). Tests and Measurement: A Tale Bearer or True Witness? *Inaugural lecture series 253*.Obafemi Awolowo University, Ile-Ife. Nigeria.

Cizek, G. (2001).*An overview of issues concerning cheating on large-scale tests.* Paper presented at the annual meeting of the National Council on Measurement in Education. Seattle

Drasgow, F., Nye, C. D., Guo, J., &Tay, L. (2009). Cheating on proctored tests: The other side of the unproctored debate. *Industrial and Organizational Psychology*, 2, 46-48.

Idika, D. (2015). Parents' concern about the use of Computer-Based Testing (CBT) for UTME in Cross River State. *Nigerian Journal of Educational Research and Evaluation.* 14 (3), 1-9.

Idika, D.,Shogbesan, Y. O. & Ogunsakin, I. B. (2016). Effect of test Item Compromise and test Item practice on the Validity of Economics Achievement Tests scores among secondary school students in Cross-River State. *African Journal of Theory and Practice of Educational Assessment* (AJTPEA). 4: 33-47.

Jurich, D. P., DeMars, C. E. & Goodman, J. T.(2012). Investigating the Impact of Compromised Anchor Items on IRT Equating Under the Nonequivalent Anchor Test Design. *Applied Psychological Measurement.* 36: 291-308, doi:10.1177/0146621612445575.

Lievens, F., &Burke, E. (2011). Dealing with threats inherent in unproctored Internet testing of cognitive ability: Resultsfrom a large-scale operational test program. *Journal of Occupational and Organizational Psychology*, 84,817-824.

McLeod, L. D., Schnipke, D. L. (1999). *Detecting Items That Have Been Memorized in the Computerized Adaptive Testing Environment.*Paper presented at the Annual Meeting of the National Council on Measurement in Education. Montreal, Quebec, Canada.

Ojerinde, D. (2015). Innovations in Asssessment: JAMB Experience. *Nigerian Journal of Educational Research and Evaluation.* 14 (3), 1-9.

Royal, K. D., & Puffer, J. C. (2012).Cheating: its implications for ABFM examinees. American Board of Family Medicine.10(3) 274-275. doi: 10.1370/afm.1408.

Shogbesan, Y. O. (2021). *Sensitivity of Economics Multiple-choice Item Parameters to Item Compromise among Secondary School Students in Ogun State, Nigeria.* An unpublished PhD. Thesis, Faculty of Education, Obafemi Awolowo University, Ile- Ife, Nigeria.

Shu Z, Henson R, Luecht R. (2013). Using Deterministic, Gated Item Response Theory Model to detect Test cheating due to Item Compromise. *Psychometrika*. 78 (3):481–97.

Tambawal, M. U. (2013).Examination Malpractices, Causes, Effects and Solutions**.** *Being a paper presented at the stake holders forum on raising integrity in the conduct of examinations in the Nigerian educational system on Thursday, 7th February 2013.*

Wollack, J. A., & Cizek, G. J., (2017). Security Issues in Professional Certification/licensure Testing. In S. Davis-Becker, & C. W. Burkendahl (2017): *Testing in the professions: credentialing policies and practices*. (New York): Routledge.

Zara, A. & Pearson, V. (2006). *Defining Item Compromise.*Paper presented at the 2006 annual meeting of the National Council on Measurement in Education, San Francisco.

Zimmermann, S., Klusmann, D., Hampe, W. (2016). Are Exam Questions Known in Advance? Using Local Dependence to Detect Cheating. *PLoS ONE*, 11(12): e0167545. https://doi.org/10.1371/journal.pone.0167545

Zumbo, B. D., &Hubley, A. M. (1998). *Differential item functioning (DIF) analysis of a synthetic CFAT*. [Technical Note 98-4, Personnel Research Team], Ottawa ON: Department of National Defense.