ITEM PARAMETER DRIFT IN CERTIFICATE EXAMINATIONS AND IT'S IMPLICATION ON DECISION MAKING

Dr (Mrs) Matilda Uvie Orheruata

Abstract

In certificate examinations, the validity of inferences made from the examination scores are of great importance. To make valid inferences regarding the examinee's ability, the examination scores must be true representative of the examinees knowledge. When examination scores fail to represent examinee's ability, it creates a misrepresentation that could threaten the fairness and validity of the examination scores. One phenomenon that is capable of causing such score misrepresentation is item parameter drift (IPD). This paper reviews the concept of IPD, causes and consequences of drift. It throws light on some methods for detecting IPD along with its implication on decision making with regards to judgement after detecting drift.

Keywords: Item parameter, Drift, Certificate examinations, Implication, Decision making

Introduction

Assessment outcomes provide valuable information that may be used to improve educational activities and as such the appropriate development and use of these assessments are essential requirements for responsible professional practices in educational testing and measurement. For examination bodies that administer multiple test forms overtime, it is critical to maintain a stable reporting scale so that scores can be comparable across years, administration and test forms.

Large-scale examination bodies that conduct norm-referenced standardized test have new items developed or old items revised from previous measures (items pool) in order to administer instruments that are more valid, reliable, sensitive and interpretable. The item pool of these examination bodies consist of a set of items in which the item parameters (difficulty level and discrimination index) have been calibrated. However, as pointed out by Demars (2004), the item parameter in large- scale examinations could become less theoretically stable especially with testing programmes that rely on a large bank of items to select from when building assessments. Though good quality items may

be selected and secured carefully by large- scale examination bodies but systematic influences may cause the item parameters estimates to change or become theoretically unstable over time. To the extent that the item becomes easier and less discriminating, causing errors in proficiency estimation using the original item parameters. This change in item parameters over time is referred to as item parameter drift (IPD) (Goldstein, 1983; Bock, Muraki & Pfeiffenberger, 1989).

Item Parameters Drift (IPD)

Item parameter drift is simply the shift of item parameter estimates from the acceptable theoretical scale. In other words, the concept of item parameter drift in measurement holds when there is a violation in the stability of the parameter scale. Item parameter drift is a phenomenon that examinees comparability of violated items across time or testing occasions (Orheruata, 2015).

In the 1980s researchers introduced the concept of IPD to represent the changes in item parameters over time (Goldstein, 1983). If parameter values Fluctuate more than would be expected due to measurement error alone, then it can no longer be assumed that parameter values are invariant over testing occasions. This violation may diminish the utility of an item over time and may warrant removal from the item bank (Clark, 2013). One basic way of finding out the IPD of a test is to try the items out and analyze their behaviour empirically. Two fundamental indicators for making this analysis are the difficulty level and the discrimination power (Sim & Rasaiah, 2006). In reality it is typical to expect IPD over multiple testing occasions (Wollack, Sung & Kang, 2006). By and large, naturally occurring amounts and magnitude of drift tend to have a very minor impact on the ability distribution of examinees.

Item parameter drift might be observed in test items in a number of ways. According to Michaelides (2010) and Clark (2013), Item difficulty values may fluctuate over administrations, with items becoming easier or hard overtime. Item discrimination values might also vary over administrations, parameter drift will affect item difficulty estimates to a stronger degree than item discrimination estimates. Parameter drift may be observed on both paper- and- pencil test and computer adaptive test forms. Similarly, fluctuations in item parameters may also be observed in classical test theory (CTT) but are found to be commonly associated with item response theory (IRT) applications Thus, regardless of the testing programme, when items are repeatedly administered consideration should be given to the occurrence of parameter drift. Also, in instances where parameter drift is observed, measurement professionals should be mindful of the features of the item and test administration that could have potentially elicited the change in parameter estimates overtime (Sukin, 2010 & Clark, 2013).

Causes of Item Parameter Drift

Item parameters may be expected to change overtime due to random error but majorly, parameter drift may be attributed to systematic changes that explain the difference in parameter estimates overtime. In 1980s when the concept of IPD was introduced, researchers found that one potential source of IPD was content and context effect such as change in content and curriculum coverage, mass media,lack of item pool maintenance, increase in teaching and exercise, Immense teaching-to-test, item exposure, item position or location, security breaches, test preparation and historic events could change how an item originally performs. Test items may also display IPD because of excessive item exposure or poor control of security (Mislevey, 1982; Goldstein, 1983 and Bock et al., 1998).

The changes in content and curriculum being covered by examinees can cause a drift in item parameters overtime. Should a change be made to content standards or a new curriculum adopted the content that items assess may receive more or less coverage, leading to differences in parameter estimates overtime. For example, Mislevey (1982) proposed a five-step procedure to account for item parameter drift. An example was a fourth-grade science test item about the metric system, which was found to be closely associated with the coverage of instruction. The time teachers spent in teaching the metric system was longer than that spent in teaching the English system, which resulted in declining difficulty for items concerning the metric system but increasing difficulty for the English system items. Any time there is a shift in the curriculum or content being covered; parameter drift might be expected to occur. Also, Bock *et al*., (1989) found differential linear drift of the item location parameters in items of a college board Physics achievement test over 10 years. They associated the direction of the drift with the content of the items in a pattern that reflected a changing emphasis in secondary school physics curricula. Among 29 items, 21 were flagged for evidence of parameter drift. Of these, items, 11 became differentially easier overtime while 10 became differentially harder. The change in difficulty was attributed to a change in the focus of the physics curriculum across that span of time. Similarly, Sykes and Fitpatrick (1992) investigated the stability of item parameter estimates with a large number of items from consecutive administration of a professional licensure examination into four content categories' they detected a significantly greater drift of Rasch difficulty parameter estimates. They attributed the differential changes in item difficulty value to shift in curricula emphasis.

Similarly, when students' classroom achievement is measured according to the curriculum, the test is initially developed based on the curriculum. Once test administration begins, however, lessons taught in the classroom tend to receive more weight in items that appear in the test, which helps students earn better test scores. Eventually, this could significantly change the actual class curriculum and as a result, test items containing content that was heavily weighted in the classroom could become easier

for test takers to handle than they originally were. Also, content that is emphasized in the mass media may also influence general knowledge on particular topics, causing some items to appear less challenging or less discriminating overtime. Another common cause for changes in item parameters overtime is threats to test security. In stances of parameter drift can be caused by examinee disclosure of items on the assessment and other forms of test fraud, which may be particularly common in large- scale examinations. Though, a security breach could be one cause of IPD, but only rarely; most testing programs take test security issues very seriously and make every effort to protect test items from illegal public exposure (Lavelle, 2008).

Maintaining an item bank properly, however, is a challenging task. Bock et al., (1989) reported that, since the quality of assessments depends heavily on the quality of item banks, proper maintenance of item banks over time is therefore critical. If an item bank is not monitored for drift over years, it is likely that the percentage of drifting items as well as the magnitude of the drift may accumulate over time and have detrimental effects on the measurement of intended construct. Over time, test items in an item bank are often re-used. Each re-use increases the likelihood that an item is improperly exposed and made available to test takers prior to testing day. Even if there were no direct threat to test security, changes in the interaction between a test item and a test taker still could occur over time for a variety of reasons.

Training in test-wise strategies can also cause a change in parameter estimates overtime. This is a source of IPD that is much harder to control. As noted by Guo and Han (2011) and Michaelides (2010) that, practicing with related test problems is a legitimate learning technique and is often encouraged but some examinees focus too much time and effort on test-taking strategies rather than on the skills and knowledge that the test will measure. As a result, some test items may become easier to test takers who practiced specific types of test items simply due to familiarity with the item and not necessarily because they improved their proficiency in the tested skill. This source of IPD can be a serious threat to test validity; such test takers' attempts are not uncommon especially when the test is high stake. Item over-exposure may also result in parameter drift, as students come to expect certain items will be included on the assessment and prepare accordingly. Higher-level changes overtime can also prompt a change in item parameters. Instances where the population changes dramatically across test administrations can be one such cause of parameter drift.

Changes made to the item between administrations may also cause parameter drift. This might include a change in the scoring of the item, mode of presentation, position on the form, or formatting and other modifications to the item's presentation. Similarly, administering a pre- and post -test or practice and scored test may result in drift between administrations, particularly due to changes in motivation or practice effects resulting from being administered the same items repeatedly over a short period of time. For example, an analysis of drift for a national computerized adaptive exam revealed

between 32 and 49% of items were flagged for drift when shifting between pre-test and operational use over a five-year period (Bergstrom, Stahl &Netzky, 2001).

A historic event is another possible cause of IPD. For example, a national presidential election can raise the public's political awareness, and hence could increase test takers' familiarity with politically related content that might appear in a test item. Test items with content that is sensitive to historic events are relatively easy to identify, however, and usually are excluded during item pool construction. As a result, IPD due to historic events usually is inconsequential (Guo& Han, 2011).

Consequences of Item Parameter Drift

The presence of Item Parameter Drift (IPD) poses a threat to the fairness and validity of test scores, thereby introducing trait- irrelevant differences that impact performance on the item over time. This threat could amount to some aftereffect that will jeopardize the fair interpretation of test scores from year to year. When sizeable magnitude of IPD exists in an achievement instrument, the amount of measurement error in scores produced by that instrument increases, thereby leading to reduction in test reliability. This in turn increases the potential for misclassifying candidates whose true scores fall at or near the passing score. Item parameter drift can have impact on examinees classification accuracy. This may lead to false decision in certificate examination. Test scores are compared over time, failing to identify drift could complex the comparison being made of examinees performance over time and this could disadvantage individual examinees.

Shift in parameter values can complicate the diagnosis of mastery of specific skills, due to items appearing differentially easy or hard over time (Clark, 2013). Similarly, anchor items that exhibit item difficulty and/or discrimination parameter drift, the resulting equating coefficients will also be influenced by the change in the item parameters and as such introduce equating error into the equating of test forms. Inclusion of anchor items exhibiting item difficulty drift will impact the passing rates (Huang &Shyu, 2003; Miller & Fitzpatrick, 2009).

Methods of detecting Item Parameter Drift

Changes in parameter values for different subgroups have been referred to as differential item functioning (DIF) while changes across testing time have been referred to as item parameter drift (IPD).**In light of the conceptual similarities between** item parameter drift (IPD) and *differential item functioning*(DIF), many of the methods employed to measure DIF within a test are also applied to measure IPD across test forms. Literature have revealed that Identifying DIF within a test involve a number of methods that are also use for detecting IPD (Sykes & Ito, 1993; Kelkar, Wightman &Leucht, 2000 &DeMars, 2004). Rather than comparing sub-groups (DIF), parameter estimates are compared

across time points or across administrations to determine whether significant difference is found between the values. The Chi-square related methods have been found to be commonly used for detecting IPD. In these methods, an item is identified to function differentially if for all person's of equal ability (that is, equal to the total score on a test containing the item) the probability of a correct response is the same regardless of each person's group membership. The Chi-square related methods Include, Mantel–Haenszel,Scheuneman's Modified, Lord's chi-square test amongst others.

The Mantel–Haenszel (M-H), a nonparametric approach for identifying DIF has been successfully used to detect drift (Michalides, 2008; Guerl, Jordan & Ackerman, 2000; Wei & Meyer, 2013). The Mantel–Haenszel test statistic according to Wiberg (2007) is based on the odd ratio between correct and incorrect responses, between a reference and a focal group when conditioning on total test scores. **M-H have been** found an efficient statistical method but cannot detect IPD for more than two groups (Kin, Cohen & Park, 1995; DeMars, 2004). Also the stability of the estimates of odds ratio in each score group may be affected by small samples. The z-statistics for the exact unsigned area measure has also been found to be effective for detecting drift (Jones & Smith, 2006; Sukin, 2010). The Scheuneman's Modified Chi- square (Scheuneman, 1987) is another method that compares various groups based on ability level on the basis of observed total test scores. With this method, an item is identified to exhibit DIF, if for all persons of equal ability, the probability of a correct response is the same regardless of group membership. The Lord's chi-square test for detecting DIF has been used by Donoghue and Isham (1998); Kim and Nering (2007); Wei and Meyers (2013) to successfully identified items exhibiting drift.

Parameter drift can also detected by methods based on Item Response Theory (IRT). These methods describe the relationship between an examinee's ability level and the probability of answering an item correctly. IRT models have been successfully used to determine if differences exist in parameter estimate across two or more administrations. The versions of the IRT models include the one- parameter, two-parameters and the three- parameters logistic models. Of the three versions, the three- parameters logistic models was found to produce more robust and detailed information on drift under standard test conditions. The three- parameters a logistic model was successfully employed by Orheruata (2015) for identifying drift in 2012 to 2014 West African Examinations Council (WAEC) and National Examinations Council (NECO) Senior School Certificate Examinations in Agricultural science objective test. Literature revealed that the one-parameter and the two - parameters logistic models are have been commonly used to determine drift (e.gSykes and Fitpatrick, 1992;DeMars, 2004; Jones and Smith, 2006) for simplified interpretations.

The IRT models have found to be the most direct and sensitive methods of determining drift. They provide more information regarding psychometric properties of individual assessment items. Graphical representation based on IRT item characteristic curve (ICC)

has also been adapted for identifying drift. The item characteristic curve for each item links the probability of correctly answering the item to examinees ability. Specifically, an ICC plots the probability of responding to an item as a function of the latent trait underlying performance on the item of the test. ICC plots provide useful visual representation of changes in parameters estimates (Wollack, Cohen & Wells, 2003). The Logistic regression (LR) method is a well-known statistical procedure proposed by Swaminathan and Rogers (1990). LR is for detecting DIF and has also been used for detecting drift for two or more administrations (Amery, Zheng, Siok& Bruno, 2008;Jodan&Guerl, 2001). According to Tabachnick and Fidell (2000). LR is based on modeling the probability of answering an item correctly by group membership and a conditioning variable, usually the observed total test score. A large weight of evidence so far supports the use of these DIF methods if differences exist in item parameter estimates across test occasions and as such may be regarded as preferred methods for identifying IPD for certificate examinations.

Implication of IPD on Decision Making

The implication is viewed as the judgemental decisions, which should be taken with regards to the use of drifted items. The premise and justification of a standardized test is that its item parameters must be stable over time. A violation of this premise is referred to IPD. Though, IPD is a necessary occurrence in practice but when severe magnitudes of drift exist in a certificate examination instrument, it becomes a concern. This is because drift directly impact on the performance of examinees. Once drift is detected, the appropriate judgemental procedures may be indispensable to determine whether or not the drifted items should be kept or removed from rotation.

The statistical finding of IPD may not necessarily warrant the removal of items that are identified as drift, rather it is necessary to apply a follow-up analysis (e.g content or context analysis). This is because the occurrence of drift is linked to potential source such as content and context effect. It is a matter of policy as to what should be done when an item is identified as displaying parameter drift. However, items identifying as displaying parameter drift could be targeted for review by content experts. Items could be kept or discarded from the item pool based on the judgement of the content specialists and test developers. Each examination bodies has policies in place that specified that items should be discarded should be discarded when certain amount of parameter drift is identified.

It is therefore advisable, for examination bodies to periodically determine parameter drift of their examination items in order to drastically reduce drift especially if drift is unidirectional. Also, sources of drift should be considered and addressed to possibly

block future occurrence.

References

Amery, D.W., Zhen, L., Siok, L.N.,& Bruno, D.Z. (2008). Investigating and comparing the item Parameter Drift in Mathematics anchor items in TIMSS between Singapore and U.S.A. *TIMSS Technical Report*. 1(6) 59-64.

Bock, R.D, Muraki, E.,&Pfeiffenberger, W. (1998) Item pool maintenance in the presence of item parameter Drift. *Journal of Educational Measurement*, 25 (94), 275-278.

Bergstrom, B., Stahl, J.A. &Netzky, B. A. (2001). Factors that influence item parameter drift. Paper presented at the annual meeting of the American Educational Research Association, Seattle.

Clark, A (2013). Review of parameter drift methodology and implications for operational testing. Paper presented at the Annual Meeting of the America Education Research Association, San Diego. June 13th2013.

DeMars, C.E. (2004). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education, 17*(3),265–300.

Donoghue, J. R, &Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement, 22(1)*33-51.

Gierl, M. J, Jodoin, M. G., & Ackerman, T. A. (2000, March). *Performance of Mantel-Haenszel, simultaneous item bias test, and logistic regression when the proportion of DIF items is large.*Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement*, *20*, 369–377.

Guo, F.,& Han, K.T. (2011). Potential impact of IPD due to practice and curriculum change in item calibration in CAT. *GMAC Research Report* RR 11-02.

Guo, F.,& Wang, L. (2005). Evaluating scale stability of a computer adaptive testing system. *GMAC Research Report*, RR-05-12.

Hambleton, R K. (1989). *Principles and selected applications of item response theory.* Educational Measurement, (3rd ed.),pp. 144-200. New York, Macmillan.

Huang, C.,&Shyu, C. (2003, April). *The impact of item Parameter drift On equating*. Paper presented At the Annual meeting Of the National Council on Measurement In Education, Chicago, IL.

Jodoin, M. G.,&Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection.*Applied Measurement in Education, 14*(4),329-349.

Jones, P. E., & Smith, R. W. (2006). *Item parameter drift In certification Exams and its impact on pass---fail decision making*. Paper presented At the annual meeting Of the National Council on Measurement in Education, San Francisco, CA. April 20th 2006.

Lavelle, L. (2008). Shutting down a GMAT cheat sheet. *BusinessWeek*.

Kelkar, V., Wightman, L. F., &Luecht, R. M. (2000, April). *Evaluation of the IRT parameter invariance property for the MCAT.*Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.

Kim, W.,&Nering, M. (2007). *Evaluation of equating items using DFIT.* Paper presented at the meeting of the National Council on Measurement in Education, Chicago, IL.

Michaelides, P.M. (2010). A review of the effects on IRT item parameter estimates with focus on misbehaving common items in test equating. *Journal on Frontier in Psychology,* 1, 167-171.

Mislevey, R.J. (1982, April). *Five steps toward controlling item parameter drift*. Paper presented at the annual meeting of the American Educational Research Association, New York.

Orheruata, M. U (2015). Item parameter drift of 2012 to 2015 WAEC and NECO SSCE Agricultural Science multiple choice items using item response theory. P.hD Dissertation. University of Benin, Benin City, Nigeria.

Scheuneman, J. D. (1987). An empirical exploratory study of causes of bias in test items. *Journal of Educational Measurement*, 24(2),99-118.

Sim, S.,&Resiah, R. I (2006). Relationship between item difficulty and discrimination indices in True/False types of multiple choice questions. *Academic Multidisciplinary Journal Singapore* 3(5) 67-71.

Sukin, T. A. (2010). Item parameter drift as an indication of differential opportunity to learning: An explanation of item flagging methods and accurate classification examiners Retrieved from:htt:www.ScholarworkU mass.edu/openness-dissertation.

Skykes, R. C.,& Ito, K. (1993, April). *Item parameter drift in IRT-based licensure examinations.* Paper presented at the Annual Meeting of the National Council on Measurement in Education. Atlanta, GA.

Tabachnick, B.G.,&Fidell, L.S. (2001). *Using multivariate statistics:* Boston: Allyn and Bacon

Wiberg, M. (2007). Measuring and detecting differential item functioning in criterion-referenced licensing test. A theoretical comparison of methods.*Retrieved from:*www.edusci.umu.se/digitalAssets/59/59534_em-no-60.

Wei, X. & Meyers, J. P. (2013, April). *Evaluation Of four robust z procedures for detecting Item parameter Drift in The 3PLM.* Paper presented At the Annual meeting Of the National Council of Measurement in Education, San Francisco, CA.

Wollack, J A, Cohen, A. S & Wells, C. S. (2003). A method for maintaining Scale stability in the presence of test speededness. *Journal of Educational Measurement, 40*,307-330.

Wollack, J.A, Sung, H. J & Kang, T. (2006). *The impact of compounding item parameter drift on ability estimation*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco. 12th- 16th July.