Educational Measurement
13(51), Spring 2023, 124-163
jem.atu.ac.ir
DOI: 10.22054/JEM.2021.38853.1882

ATU PRESS

Application of Two-Parameter Nested Logit Model in Identifying the Source of DIF in Multiple-Choice Items

Hassan Moshtaghian Abargouei — Ph.D. Student in Deliberation and Measuring, Allameh Tabataba'i University, Tehran, Iran. E-mail: kavir311@gmail.com

Mohammad Reza Flasafi Nejad* — *Corresponding Author*, Associate Professor, Department of Deliberation and Measuring, Allameh Tabataba'i University, Tehran, Iran. E-mail: falsafinejad@yahoo.co.uk

Ali Delavar — Professor, Department of Deliberation and Measuring, University of Tehran, Tehran, Iran. E-mail: dr.delavarali@gmail.com

Noor Ali Farrokhi — Professor, M Department of Deliberation and Measuring, Allameh Tabataba'i University, Tehran, Iran. E-mail: farrokhinoorali@gmail.com

Abstract

Identifying distractors as sources of Differential Item Functioning(DIF) in polyotomous items has great importance to designers and analysts. Although DIF is one of the common methods for examining the measurement invariance, It is accompanied by challenges and limitations, especially in multiple choice items. The purpose of this study was to assess the performance of Nested logit Model(NLM) for detecting Differential Distractor Functioning(DDF) by using experimental (simulated data) and descriptive-analytical (real data) methods. Six items were simulated under different conditions of difficulty and slope, ability distribution, presence or absence of DIF/DDF, and DIF/DDF magnitude, with a sample size of 2000 and 50 replicates. The data of Math Entrance Exam (D-form,2018), with a random sample of 2000 men and women constituted the real data. Based on the results of the simulation analysis: The NLM revealed 88% of DIF and 97% of DDF, on average. the Type I error rates is very close to the theoretical expected values, although it showed some inflation in unequal distribution conditions. according to the findings, the detection rate was influenced by the item parameters(difficulty and slope) and the DIF or DDF levels. Based on real data analysis, 2 items represented both DIF(Large and Medium) and DDF (Partial to Moderate) simultaneously, whereas in the NRM

How to Cite: Moshtaghian Abargouei, H., Falsafi Nejad, M. R., Delavar, A., & Farrokhi, N. A. (2023). Application of Two-Parameter Nested Logit Model in Identifying the Source of DIF in Multiple-Choice Items. *Quarterly of Educational Measurement*, *13*(51), 124-163. doi: 10.22054/JEM.2021.38853.1882

Original Reserch
Received: Dec. 25, 2018
Accepted: Dec. 08, 2021
ISSN: 2252-004x
eISSN: 2470-6240

approach, 11 items detected as DIF/DDF; so, as expected the approaches based on "divided by distractor" strategy, fewer items were detected as DIF/DDF. The NLM approach, while separating the DDF from the DIF test, allows for a clear evaluation of whether the distractor may be responsible for DIF. Since high-stakes tests have a special role in selection and DIF and DDF analyzes have a special place in determining the validity and measurement invariance of these exam items, it is recommended to screen the bias items, DIF/DDF comprehensive analyzes based on NLM be used.

Keywords: Differential item functioning, Differential distractor functioning, Two-parameter nested logist model, Simulation studies

1. Introduction

DIF analysis is essential for ensuring test fairness and maintaining measurement validity. Despite the focus on dichotomous-scored items, multiple-choice items are also subject to DIF, with distractors being a frequent culprit (e.g., Banks, 2009; Penfield, 2008, 2010; Suh & Bolt, 2011; Suh & Talley, 2015). As mentioned previously, the concept of DDF is often employed to explain instances where individual distractors display distinct levels of attractiveness among individuals of the same ability but from different groups, as per the definitions defined by Suh & Bolt (2011). As pointed out earlier, DDF in itself is not problematic; however, when DDF is related to DIF, it becomes an issue, as one group may gravitate towards particular distractors, decreasing their chances of attaining the right answer (Banks, 2009). Studying DDF in conjunction with DIF analysis is crucial to gaining insight into the reasons behind DIF in multiple-choice items. As per Suh and Bolt (2011), there has been an apparent lack of consistency in the terminology employed in the psychometric literature to identify DDF effects.

The aforementioned conceptualization of DDF can vary, wherein the term is alternatively used to denote a differential level of performance between distractors or the correct response ("divide-by-distractors" and "divide-by-total" frameworks, respectively). Considering your referenced conceptualizations, the application of the "divide-by-distractors" approach may better address the potential connection between distractors and DIF in the correct response. Identifying the distractor causes of DIF can indeed offer insightful information that can be leveraged for making productive changes to item adjustment and design, thereby contributing to enhanced fairness and reliability in

assessment. Despite several methods being developed for the identification of DDF, the specific advantages of these techniques and their relative efficacy under varying conditions remain undecided. The limited research on the subject has yielded inconsistent results, as showcased by the work of Su and Talley (2015).

The 2PL-NLM approach, as outlined by Suh and Bolt (2011), is centered on identifying DDF based on incorrect responses only, adhering to the concept of "divide-by-distractors." Given the aforementioned DDF conception, where DIF affects the correct response. The occurrence of DDF may be absent. Therefore, these approaches enable the separation of the two DIB sources via a two-step process, where a DIF analysis is done first and subsequently followed by a DDF analysis to assess if this distractor causes the inaccurate appearance of DIF on the correct answer option. Although the nested logit approach based on the "divide-by-distractors" framework appears to provide multiple benefits in identifying the potential sources of DIF (Su and Bolt, 2011), the currently available evidence is insufficient to validate this assertion. The present study sought to evaluate the efficacy of the two-paramater nested logit model (2PL-NLM) in identifying differential distractor functioning in diverse circumstances, utilizing Monte Carlo (MC) analyses of simulated data and inspection of actual data sets.

Research Question(s)

1. Does the ability distribution of groups (same and different), the difficulty and discrimination of items, the size of DDF (medium, large), and the size of DIF (small, medium, large) placement affect the identification of the differential functioning of a question using the 2-paramater nested logit model?
2. What is the correlation between differential item functioning and differential distractor functioning in actual tests?

2. Literature Review

Several researchers are examining distractor choices in tests. For example, Green et al. (1989) used log-linear models to check whether there is a potential interaction between a group attribute (e.g. gender) and distractor choice when ability is kept constant. Likewise, the second approach investigates DDF by considering the conditional probability of each distractor when group attributes such as gender are held

constant (Dorans et al., 1992). Additionally, a third method draws on item response theory (IRT) and the likelihood ratio test (LR) to measure DDF (Thissen, Steinberg, & Gerrard, 1986; Thissen, Steinberg, & Wainer, 1988, 1993). As an example, Thissen et al. (1993) applied an LR test to gauge the differences between groups with respect to the response curves associated with all response categories, as per the Multiple-Choice Model (Thissen, Steinberg, & Fitzpatrick, 1989).
Indeed, Suh and Bolt (2011) proposed a two-step LR test approach under the 2PL-NLM and compared it with an LR test conducted based on the Nominal Response Model (NRM; Bock, 1972) as a complementary alternative. Additionally, Penfield (2008, 2010) put forth an odds ratio estimate approach under the Nominal Response Model (NRM) to enhance DDF detection power and appraise various ways in which distractors potentially lead to DIF.

3. Methodology

To evaluate the effectiveness of distractors as potential contributors to DIF in practice, a combined simulation- and actual-data analysis experiment was initiated.

Detection Method

According to Suh and Bolt (2011), the 2PL-nested logit model (2PL-NLM) has been suggested as a viable alternative to conventional multinomial logistic models (such as the Nominal Response Model NRM; Suh & Bolt, 2011) and claimed to offer an advantageous approach to probing DIF and DDF in multiple-choice items (Suh & Bolt, 2011).
Under the 2PL-NLM, the probability that an examinee of ability θj chooses the correct response category on item i is modeled as the traditional two-parameter logistic (2PL) model and given by:

$$p_i(q_j) = \frac{1}{1+ \exp^{-(b_i + a_i q_j)}}$$

Where bi denotes an intercept parameter and αi is a slope parameter for item i. The probability that the examinee selects distractor category v ($v = 1, 2, \ldots, m$) is modeled as the product of the probability of an incorrect response and the probability of selecting distractor category v conditional upon an incorrect response:

$$p_{iv}(q_j) = \left[1 - \frac{1}{1+\exp^{-(b_i + a_i q_j)}}\right] \left[\frac{\exp^{z_{iv} + l_{iv} q_j}}{\sum_{k=1}^{m} \exp^{z_{iv} + l_{iv} q_j}}\right]$$

where ζs and λs are the intercept and slope parameters for the distractor categories, respectively. Such a nested logit modeling framework can be appealing for studying DDF, as this framework creates a separation between the correct response category parameters and distractor category parameters that can be used to evaluate DDF independent of DIF (Suh & Bolt, 2011).

Indeed, Suh and Bolt (2011) laid out a three-tiered framework under the 2PL-NLM with varying constraints for investigating DIF and DDF: (a) a "compact model", whereby all item parameters of the under-investigation item are set equal across groups; (b) an "augmented model", in which the distractors' parameters, specifically, are constrained to be equal across groups; and (c) a "second augmented model" wherein none of the studied item's parameters are constrained to equality."

The LR test statistic is calculated as $G_2 = -2 \log L_1 - (-2 \log L_2)$, where $\log L_1$ and $\log L_2$ are the log likelihoods for a simpler model and a more complex model, respectively, and is distributed as a χ^2 with degrees of freedom (*df)* equal to the difference in the number of parameters estimated between the two models. Indeed, a DIF test can be accomplished with the calculation of the G2 statistic in two steps. First, by comparing the Compact and the First Augmented models (TEST1). Then, a G2 statistic test of the First and Second Augmented models (TEST2) can assess the presence of differential distractor functioning (DDF). When this investigated item yields significant results for both TEST1 and TEST2, indicating DIF and DDF, the distractors seem to influence the DIF at least in part.

Data Simulation

The performance of the likelihood ratio (LR) test in detecting DDF, in the 2PL-nested model, was evaluated on three datasets: (1) "Non-DIF" data without either DIF or DDF for the investigated items; (2) "DIF" data with only DIF, implying the absence of a distractor influence; and (3) "DIF+DDF" data, signifying the combined presence of both DIF and DDF. The combination of conditions allows one to probe the Type

I error rates and rejection rates (primary outcomes) associated with the likelihood ratio (LR) test when implemented via the multi-group 2PL-nested model. The simulated data collection tool was a test consisting of 36 four-choice items. We designed it via the mcIRT package in the R environment. We constructed the test for 1000 focal group and 1000 reference group participants, employing 50 replication cycles.

The selected independent variables include: 1. Question parameters (e.g., difficulty, discrimination); 2. Population ability (or ability distribution); 3. Differential functioning (or DDF) (present/absent); 4. Magnitude of the DDF; 5. Sample size; 6. Test length. These factors were the same for both reference groups.

Ability distribution: Differences in ability distribution can affect the detection of DIF (Judin & Girel, 2001); Therefore, in order to simulate the situation in which the ability level of the groups is the same (no effect) from the standard distribution $\theta \sim$ Normal(0, 1) for both groups and to simulate the situation in which one group has more ability than the other group (presence of effect), the $\theta \sim$Normal(0, 1) distribution was used for the reference group and the $\theta \sim$ Normal($-.5$, 1) distribution was used for the focal group.

Anchor items sets: the same set of anchor items was used across all three conditions. 2PL-nested model parameters for 30 anchor items with four-response categories (i.e., one correct answer and three distractors) were generated using the following distributions: $\alpha \sim$ uniform (.75, 2) and $\beta \sim$ uniform (-2.5, 2.5) for the correct response category, and $\lambda \sim$ uniform (-2, 2) and $\zeta \sim$ uniform (-2, 2) for the distractor categories. The constraints were imposed for the distractor categories.

Non-DIF data: six studied items without DIF were simulated by crossing two slope parameter values, α = 1.25 and .75, with three intercept parameter values, β = 1.5, 0, and -1.5, for the correct response category. The item parameters for distractors were λ_v =0.26, -0.28, 0.02 for slope and ζ_v = -0.22, -0.14, 0.36 for intercept and were fixed across all six items.

DIF data: for studied items in the DIF data sets, three DIF levels were simulated: low (Δb = 0.25), medium (Δb = 0.5), and high (Δb = 1). Also, DIF in the slope parameter of item was Δa =0.3. For the reference group, the Non-DIF item parameters for the six studied items were consistently used, whereas for the focal group, .25, .5, and 1.0 were

subtracted from the Non-DIF β parameters, making the items more difficult. These values represent low (DIF-L), medium (DIF-M), and high (DIF-H) levels of DIF. DIF in the slope parameter was introduced at a shift level of .3, such that the slope parameters for the focal group were set .3 higher than for the reference group. Thus, three different levels of DIF in the correct response category for each of the six studied items were generated, resulting in a total of 18 DIF items simulated. The distractor category parameters for both groups were set equal to the parameters used to generate the Non-DIF data sets.

DIF+DDF data: for items with DDF two DDF levels: medium ($\Delta\zeta$ = 0.4) and high ($\Delta\zeta$ = 1.2) were simulated. DDF in the slope parameter of distractors was $\Delta\lambda$ = 0.3 that were introduced simultaneously with DIF in the parameter of difficulty and intercept. To generate studied items for the DIF+DDF data sets, two levels of DDF related to the distractor intercept parameters were crossed with the three levels of DIF described earlier (DIF-L, DIF-M, DIF-H) to generate six different combinations of conditions. For DDF, the third category intercept (ζ_3) was arbitrarily chosen to be lower for the reference group, thus making the other two distractor intercepts higher for the reference group due to the constraint. By subtracting either .4 or 1.2 from ζ_3 for the reference group, moderate and high levels of DDF were simulated. Because these two conditions were crossed with three DIF levels, there were six different types of studied items for each condition, implying a total of 36 DIF+DDF item types.

In total, 6 items without DIF (and without DDF), 18 items with DIF, and 36 items with DIF+DDF (each with 50 replications), were generated from the combination of different conditions.

Actual Data

Actual data was obtained from "Mathematics - 30 items" sub-test, the Form D section of the specialized Science Subgroup exams of the 2018 Iranian college entrance examination. For the secondary analysis, the statistical population consisted of all students who took at least one math's test item during the exam. A sample of 1000 men (reference group) and 1000 women (focal group) was randomly selected from this population using data provided by the national measurement entity.

In order to select a suitable group of items as anchors in our data, we utilized the iterative purification technique (Lord, 1980). This process

involved conducting a likelihood ratio (LR) test to identify items that did not display significant differences between response categories, implying that they weren't exhibiting DIF. We adopted an approach initially introduced by Kim and Cohen (1995) during our analysis. After conducting the iterative purification method, we decided to discard 12 items that showed signs of DIF (significant differences in response pattern) and one item exhibiting DDF (differential distractor functioning). Thus, we selected 17 items for our analysis. We identified 17 items as anchor items, which did not display any signs of DIF or DDF during our analysis. We subsequently proceeded to examine all remaining items for those effects with the 2PL-MNL model and Bock's (1972) NRM.

4. Results

- Based on the simulation data, the nested logit model accurately identified 87.7% and 97.5% of the DIF and DDF items respectively, reflecting its strong power, especially about DDF.
- The simulation data suggests that in the absence of any differential functioning and under conditions of equal distribution of ability, type-1 error rates observed for DIF and DDF were 0.043 and 0.047, respectively. However, under conditions of unequal distribution of ability, minor increases in the type-1 error rate to 0.055 for DIF and 0.054 for DDF were noticed, which still falls within the nominal alpha range of agreement.
- Per the findings, the detection of DIF was influenced by the item's parameters, specifically its difficulty level and slope. Such that as the difficulty level increased or the slope decreased, the power of identifying DIF exhibited a rise.
- The detection rate of DDF seemed to have a high dependence on the item's parameters, specifically its difficulty and recognition capabilities. The simulation results suggested a significant rise in the detection of DDF when the item's difficulty level was lowered or its power heightened.
- The findings highlighted another noteworthy aspect the potential concurrence or absence of both DIF and DDF. The logical explanation suggested that the aforementioned separation was possible because the nested logit method incorporated the "division by deviant alternatives" approach, a framework that enables the decoupling of DIF sources with a two-step strategy, as explained by Su and Bolt (2011).

- The empirical results based on the actual data revealed multiple DIF and DDF instances. Specifically, 9 questions showcased DIF, 4 questions expressed DDF, and 2 simultaneously contained both DIF and DDF signs. Additionally, the findings revealed the superiority of the nested logit model, which relied on the "divide by deviant options" strategy, in comparison to the nominal response approach, which categorized 11 questions as having DDF. This difference indicates that the nested logit approach was more conservative in DDF detection, as noted by Su and Talley (2015).

5. Conclusion

It is certainly vital to distinguish between DIF and DDF, especially in multi-choice questions. Such insights could help in adjusting or creating alternative questions that alleviate this discrepancy, ultimately enhancing the quality of assessments. The two-step process, using the two-parameter nested logit (TPNL) model, possesses the capability to establish whether the DIF is the result of different distractor function. The separation of the DIF and DDF tests enhances the accuracy of identifying the source of the bias and providing support for content experts to identify problematic distractors. It also enables the incorporation of a robust correction mechanism into future tests.

مقاله پژوهشی

اندازه‌گیری تربیتی
دوره ۱۳، شماره ۵۱، بهار ۱۴۰۲، ۱۲۴-۱۶۳
jem.atu.ac.ir
DOI: 10.22054/JEM.2021.38853.1882

کاربرد مدل دو پارامتری لوجیت آشیانه‌ای در شناسایی منابع DIF در سؤال‌های چندگزینه‌ای

حسن مشتاقیان ابرقوئی دانشجوی دکتری سنجش و اندازه‌گیری دانشگاه علامه طباطبایی، تهران، ایران. رایانامه: kavir311@gmail.com

محمدرضا فلسفی نژاد * *نویسنده مسئول*، دانشیار گروه سنجش و اندازه گیری، دانشگاه علامه طباطبایی، تهران، ایران. رایانامه: falsafinejad@yahoo.co.uk

علی دلاور استاد، گروه سنجش و اندازه گیری، دانشگاه علامه طباطبایی، تهران، ایران. رایانامه: dr.delavarali@gmail.com

نورعلی فرخی دانشیار گروه سنجش و اندازه گیری، دانشگاه علامه طباطبایی، تهران، ایران. رایانامه: farrokhinoorali@gmail.com

تاریخ دریافت: ۱۳۹۷/۱۰/۰۴؛ تاریخ پذیرش: ۱۴۰۰/۰۹/۱۷

ISSN: 2252-004x
eISSN: 2470-6240

چکیده

مشخص کردن گزینه‌های انحرافی به‌عنوان منابع کنش افتراقی سؤال (DIF) در سؤال‌های چند ارزشی اهمیت بسزایی برای طراحان و تحلیل گران سؤال دارد. هرچند DIF روش معمول بررسی تغییرناپذیری اندازه‌گیری است؛ این رویکرد به‌خصوص در سؤال‌های چندگزینه‌ای با چالش‌ها و محدودیت‌هایی همراه است. هدف این مطالعه، ارزیابی رویکرد لوجیت آشیانه‌ای (NLM) در شناسایی سؤال‌های حاوی کنش افتراقی گزینه‌های انحرافی (DDF) با استفاده از روش تحقیق آزمایشی (داده‌های شبیه‌سازی) و روش توصیفی-تحلیلی (داده‌های واقعی) بود. ۶ سؤال، تحت شرایط مختلف دشواری و شیب، توزیع توانایی، وجود یا نبود کنش افتراقی و بزرگی DIF/DDF با نمونه‌ای به حجم ۲۰۰۰ و با ۵۰ تکرار شبیه‌سازی شد. همچنین، داده‌های فرم D آزمون ریاضی کنکور ۱۳۹۷ با نمونه‌ای تصادفی به حجم ۲۰۰۰ مرد و زن، نمونه واقعی را تشکیل می‌داد. بر اساس نتایج تحلیل داده‌های شبیه‌سازی: رویکرد لوجیت آشیانه‌ای به‌طور متوسط ۸۸ درصد سؤال‌های DIF دار و ۹۷ درصد سؤال‌های DDF دار تحت شرایط مختلف را آشکار نمود. نرخ خطای نوع اول در اغلب شرایط بسیار نزدیک به ارزش‌های مورد انتظار نظری بود هرچند در شرایط توزیع نابرابر، مقداری تورم خطا نشان داد. بر اساس یافته‌های شبیه‌سازی، نرخ تشخیص کنش افتراقی متأثر از پارامترهای سؤال (دشواری و شیب) بود و با افزایش سطح DIF و یا DDF نرخ رد افزایش می‌یافت. مبتنی بر تحلیل داده‌های واقعی، ۲ سؤال به‌طور هم‌زمان هردوی DIF (بزرگ و متوسط) و DDF (جزئی تا متوسط) را به نمایش گذاشت، درحالی‌که در رویکرد رقیبِ پاسخ اسمی، ۱۱ سؤال به‌عنوان سؤال با کنش افتراقی شناسایی شد؛ بنابراین همان‌طور که انتظار می‌رفت رویکرد NLM مبتنی بر استراتژی «تقسیم‌بر گزینه‌های انحرافی» تعداد سؤال‌های کمتری را به‌عنوان DIF / DDF دار ردگیری نمود. رویکرد دومرحله‌ای مدل لوجیت آشیانه‌ای، ضمن تفکیک آزمون DDF از DIF، امکان ارزیابی روشن از اینکه آیا گزینه‌های انحرافی مسئول احتمالی DIF هستند را میسر می‌سازد. ازآنجاکه آزمون‌های سرنوشت‌ساز نقش ویژه‌ای در گزینش افراد دارند و تحلیل‌های DIF و DDF جایگاه ویژه‌ای در تعیین اعتبار و نامتغیر بودن اندازه‌گیری سؤال‌های این آزمون‌ها دارند، پیشنهاد می‌شود جهت سرند کردن سؤال‌های سودار تحلیل‌های جامع DIF / DDF مبتنی بر رویکردهای لوجیت آشیانه‌ای مورداستفاده قرار گیرد.

کلیدواژه‌ها: کنش افتراقی سؤال، کنش افتراقی گزینه‌های انحرافی، لوجیت آشیانه‌ای دو پارامتری، مطالعات شبیه‌سازی

استناد به این مقاله: مشتاقیان ابرقوئی، حسن، فلسفی نژاد، محمدرضا، و دلاور، علی، و فرخی، نورعلی. (۱۴۰۲). کاربرد مدل دو پارامتری لوجیت آشیانه‌ای در شناسایی منابع DIF در سؤال‌های چندگزینه‌ای. *فصلنامه اندازه‌گیری تربیتی*، ۱۳(۵۱)، ۱۲۴-۱۶۳.
doi: 10.22054/JEM.2021.38853.1882

مقدمه

امروزه، منصفانه بودن[1] آزمون‌ها و تغییرناپذیری اندازه‌گیری[2] به‌ویژه در حوزه آزمون‌های سرنوشت‌ساز[3] که در مقیاس وسیع اجرا می‌شوند و بر اساس نتایج آن‌ها تصمیم‌های مهمی در مورد افراد اتخاذ می‌شود، اهمیت فراوانی پیداکرده است (Zieky, 2006). هرچند در حوزه آزمودن تعاریف بسیاری از منصف بودن ارائه‌شده است؛ اما هیچ‌گونه تعریف پذیرفته‌شده جهانی وجود ندارد. بااین‌حال، بر سر این موضوع توافق وجود دارد که انصاف در آزمون مستلزم برخورد بی‌طرفانه با همه آزمودنی‌ها در فرایند آزمودن است؛ یعنی به همه شرکت‌کنندگان در آزمون، بدون توجه به عضویت گروهی باید شانس برابری برای یک تفسیر معتبر از نمره‌های آزمون داده شود(AERA[4], APA[5], NCME[6], 2014). به‌عبارت‌دیگر، منصفانه بودن یک موضوع بنیادی اعتبار[7] است (Penfield, 2016؛ Dorans & Cook, 2016) و لذا ارزیابی عامل‌های نامرتبط–باسازه در ارتباط با پاسخ‌های نمره‌گذاری شده افراد، نقش مهمی در ارزیابی منصفانه بودن و تغییرناپذیری اندازه‌گیری دارد (Penfield, 2010). در حال حاضر روش‌های کنش افتراقی سؤال[8] (DIF) یک چارچوب عام مورداستفاده برای ارزیابی تغییرناپذیری اندازه‌گیری درزمینهٔ آزمون‌های آموزشی است (Camilli & Shepard, 1994؛ Holland & Thayer, 1988؛ Holland & Thayer, 1988؛ Penfield, 2008) و تاکنون، رویکردهای متعددی برای ردگیری DIF پیشنهادشده و مورداستفاده قرارگرفته است؛ اما یکی از محدودیت عمده روش‌های سنتی DIF آن است که تنها بر تفاوت‌های گروهی مرتبط با طبقه پاسخِ صحیحِ سؤال‌های نمره‌گذاری شده به‌صورت دو ارزشی متمرکز بوده (Suh & Talley, 2015؛ Penfield, 2010) و حتی در سؤال‌های چند ارزشی هیچ توجه ای به نقش گزینه‌های انحرافی ندارد. این مسئله، نه‌تنها موجب محدود شدن تعریف تغییرناپذیری اندازه‌گیری می‌شود (Suh & Bolt, 2011؛ Suh & Talley, 2015) بلکه، اطلاعات چندانی نیز دراین‌باره که کجا در بین گزینه‌ها اثر DIF

---

1. fairness
2 .measurement invariance
3. high-stakes tests
4. American Educational Research Association(AERA)
5. American Psychological Association(APA)
6. National Council on Measurement in Education(NCME)
7. validity
8. differential item functioning(DIF)

آشکار و تغییرناپذیری رخ می‌دهد، برای نویسندگان سؤال‌های آزمون فراهم نمی‌سازد (Penfield, 2010). به عبارت ساده‌تر،DIF تنها وجود مشکل در سؤال را مشخص می‌کند، اما اطلاعی درباره موقعیت عامل سوگیری در سؤال به دست نمی‌دهد؛ درحالی‌که برخی سرنخ‌های جالب درباره تفاوت‌های گروهی در فرآیند آزمودن را می‌توان در پاسخ‌های غلط خاصی که افراد به سؤال‌های آزمون می‌دهند، جستجو کرد (Green et al., 1989؛ Thissen & Steinberg, 1984؛ Thissen et al., 1989). محدودیت‌های ذکرشده که یک چالش اساسی در روش‌های سنتی تحلیل DIF است، سبب شده تا برخی محققان در ارزیابی تغییرناپذیری سؤال‌های چندگزینه‌ای، علاوه بر مطالعه DIF، به تفاوت‌های گروهی در احتمال شرطی مرتبط با هریک از گزینه‌های انحرافی، یعنی آنچه از آن به کنش افتراقی گزینه‌های انحرافی[1] (DDF) یاد می‌شود، نیز توجه نمایند (Green et al., 1989؛ Schmitt & Bleistein, 1987؛ Schmitt & Dorans, 1990؛ Penfield, 2010؛ Suh & Bolt, 2011؛ Suh & Talley, 2015). به‌طوری‌که امروزه در ارزیابی منصفانه بودن آزمون و تغییرناپذیری اندازه‌گیری از تحلیل جامع DDF/DIF صحبت به عمل می‌آید. بررسی الگوی اثراتِ DDF همراه بامطالعه DIF ضمن توجه به تغییرناپذیری همه سطوح گزینه‌های سؤال، می‌تواند اطلاعات ارزشمندی به‌منظور تعیین منابع DIF در سؤال‌های چندگزینه‌ای فراهم نماید و پتانسیل آن را دارد تا نقش مهمی در بررسی منصفانه بودن سؤال و آزمون ایفا کند (Bolt, 2000؛ Douglas et al., 1996؛ Schmitt, Holland, & Dorans, 1993). بااین‌وجود، هنوز اطلاعات چندانی پیرامون مزیت‌های نسبی رویکردهای مختلف DDF در دسترس نیست.

بهبود روش‌های مطالعه DDF از دهه ۱۹۸۰ صورت گرفت و تاکنون روش‌های مختلفی چون: لگاریتم خطی (Green et al., 1989) رویکرد رگرسیونی (Kato et al., 2009)؛ رویکرد مدل عاملی (Wang, 2000)؛ روش استانداردسازی[2] (Banks, 2009)؛ روش نسبت بخت تحت مدل پاسخ اسمی (Penfield, 2008)؛ رویکرد لوجیت آشیانه‌ای (Suh & Bolt, 2011) برای ارزیابی کارکرد گزینه‌های انحرافی پیشنهاد و مورداستفاده قرارگرفته است. Mapuranga و همکاران (2008) روش‌های مطالعه اثرات DDF را در چهار مقوله: الف- روش‌های مدل خطی تعمیم‌یافته؛ ب- روش‌های نمره مورد انتظار سؤال، ج-

1. differential distractor functioning(DDF)
2. SIBTEST

روش‌های نظریه سؤال پاسخ (IRT) و د- روش‌های نسبت بخت نا پارامتری دسته‌بندی کردند. رویکرد مدل خطی تعمیم‌یافته، شامل روش‌هایی مبتنی بر رگرسیون لجستیک، مدل‌های ترکیبی و مدل‌های خطی سلسله مراتبی چون مدل لگاریتم خطی است. هرچند مدل لگاریتم خطی قادر به ارزیابی اثرات هم‌زمان DIF و DDF است و اجازه یک آزمون کلی DDF را می‌دهد، اما اطلاعات مفیدی در این رابطه که کدام گزینه انحرافی مسئول معنی‌داری نتایج است، فراهم نمی‌کند (Suh & Bolt, 2011). یک محدودیت دیگر کاربرد رویکرد لگاریتم خطی آن است که شاخصی برای اثرDIF /DDF فراهم نمی‌کند، این موضوع، انعطاف‌پذیری این رویکرد را در فراهم نمودن اطلاعاتی درباره بزرگی کارکرد افتراقی مرتبط با هر طبقه پاسخ و علل اختصاصی اثراتDIF و DDF محدود می‌نماید (Penfield, 2008). از طرفی، رویکرد رگرسیون لجستیک بر پیش‌فرض‌های محکمی استوار است ازجمله اینکه نمره مشاهده‌شده X یک معرف روا از توانایی باشد و اینکه مدل رگرسیون لجستیک معرف دقیقی از روابط بین احتمال پاسخ صحیح و نمره مشاهده‌شده آزمون باشد (Camilli & Shepard, 1994)، درحالی‌که به‌خوبی معلوم است که نمره مشاهده‌شده آزمون، رابطه غیرخطی با توانایی مکنون دارد (امبرتسون و رایس، ۲۰۰۰، ترجمه شریفی، ۱۳۸۸؛ Lord, 1980). در رویکرد نمره مورد انتظار سؤال، تمرکز بر تعیین تفاوت‌های بین نسبتِ نرخ‌های پاسخ، بین گروه‌های مرجع و کانونی در هر سطح توانایی است. از مثال‌های معمول این رویکرد می‌توان به روش استانداردسازی[1] اشاره نمود که به اندازه‌گیری تفاوت بین گروهی در احتمال شرطی مرتبط با هر گزینه انحرافی می‌پردازد (Dorans et al., 1992؛ Schmitt & Bleistein, 1987). این رویکرد هرچند رفتار همه گزینه‌های پاسخ را به‌طور هم‌زمان بررسی می‌کند، اما کاملاً جنبه توصیفی دارد (Kato et al., 2009). رویکرد استانداردسازی برخلاف رویکرد لگاریتم خطی، یک اندازه اثر[2] برای DDF فراهم می‌کند؛ اما اثرات DDF به‌دست‌آمده ارتباط روشنی با اندازه اثر DDF برای مدل‌های پارامتری فراهم نمی‌کند (Penfield, 2008). به‌طورکلی آماره‌های مبتنی بر نمره آزمون در نمونه‌های مختلف یکسان نیستند و لذا در مطالعه تغییرناپذیری اندازه‌گیری رضایت‌بخش نیستند (امبرتسون و رایس، ۲۰۰۰، ترجمه شریفی، ۱۳۸۸). از دیگر مشکلات جدی روش‌های کنش افتراقی که نمره خام کل را به‌عنوان متغیر همتاسازی بکار می‌گیرد،

1. standardization approach
2. effect size

مسئله تورم خطای نوع اول است (Roussos & Stout, 1996؛ DeMars, 2010؛ Li, 2014). رویکرد سوم، مبتنی بر مدل نظریه سؤال- پاسخ (IRT) است. در این گروه، تکنیک‌های DDF از یک متغیر نهفته در تعریف فرض صفر DIF استفاده می‌کنند (Koon & Kamata, 2013). به‌عنوان‌مثال، Thissen و همکاران (1993) از یک آزمون نسبت درستنمایی[1] (LR) برای اندازه‌گیری تفاوت‌های بین گروهی در منحنی‌های پاسخ[2] مرتبط با همه طبقات پاسخ آن‌گونه که برای مدل‌های چندگزینه‌ای پارامتر سازی می‌شود، استفاده کردند. Suh & Bolt (2010) روش آزمون LR دومرحله‌ای را برای ردگیری DIF و DDF تحت مدل لوجیت دو پارامتری آشیانه‌ای معرفی کردند. درحالی‌که رویکردهای پارامتری مبتنی بر مدل‌های IRT یک چارچوب انعطاف‌پذیر و فرهیخته برای ارزیابی اثرات DDF فراهم می‌کند، بر پیش‌فرض‌های محدودکننده‌ای چون تک‌بعدی بودن، استقلال موضعی و برازش مدل مبتنی است. همچنین برآوردهای باثبات پارامترها نیازمند اندازه‌های نسبتاً بزرگ نمونه است (Penfield, 2008). آخرین طبقه از روش‌ها شامل نسبت‌های بخت[3] نا پارامتری است که تفاوت‌های ممکن در بختِ پاسخ صحیح در سطوح مختلف توانایی بین گروه‌های مرجع و کانونی را موردبررسی قرار می‌دهد. به‌عنوان‌مثال Penfield (2008) یک برآورد کننده لگاریتم نسبت بخت مبتنی بر مدل NRM را برای آزمون معنی‌داری اثر DDF پیشنهاد داد. رویکرد پنفیلد علی‌رغم ماهیت نا پارامتریک که آن را از محدودیت برازش مدل آزاد می‌کند و نیز مزیت ارائه اندازه اثر DDF برای هریک از گزینه‌ها، در ارزیابی اینکه DIF واقعاً به علت گزینه‌های انحرافی رخ می‌دهد، ممکن است کمتر مفید باشد (Suh & Bolt, 2011). به‌طورکلی، بررسی پیشینه نشان می‌دهد که در ادبیات روان‌سنجی DDF به شکل یکسانی مفهوم‌پردازی نشده است. گاهی این مفهوم به کارکرد افتراقی یک گزینه انحرافی نسبت به همه طبقات پاسخ، ازجمله طبقه پاسخ صحیح (رویکرد تقسیم‌بر کل) اشاره دارد و گاهی این مفهوم مرتبط با کارکرد افتراقی یک گزینه انحرافی نسبت به دیگر گزینه‌های انحرافی (رویکرد تقسیم‌بر گزینه‌های انحرافی) است (Suh & Bolt, 2011). مطابق مفهوم‌پردازی نخست از DDF جایی که DIF حضور دارد، DDF نیز به‌عنوان یک ضرورت آماری باید رخ دهد چراکه ممکن است یک گروه

1. likelihood ratio (LR)
2. response curves
3. odds ratio estimator

بیش‌ازحد به یک گزینه انحرافی خاص جلب شود و درنتیجه شانس آن گروه برای به دست آوردن پاسخ صحیحِ سؤال کاهش یابد (Banks, 2009)؛ بنابراین هنگامی‌که DDF به این شکل مورد تحلیل قرار می‌گیرد، اغلبِ سؤال‌هایی که DIF نشان می‌دهند، DDF را نیز به نمایش خواهند گذاشت (Middleton & Laitusis, 2007). درحالی‌که حضور هم‌زمان DIF و DDF لزوماً به معنی اینکه گزینه‌های انحرافی مسبب DIF هستند، نیست. به‌بیان‌دیگر با بکار بردن این نوع رویکرد، ممکن است نتوان به‌طور مؤثری دو منبع سوگیری در سؤال، یعنی وجود عامل سوگیری در گزینه صحیح یا گزینه‌های انحرافی را از هم تفکیک کرد (Suh & Talley, 2015). دومین رویکرد، یعنی چارچوب «تقسیم‌بر گزینه‌های انحرافی»، بر پاسخِ سؤال به‌عنوان یک فرایند سلسله مراتبی تأکید می‌ورزد و به‌موجب آن، اولین تلاش آزمودنی برای حل یک سؤال مستقل از گزینه‌های پاسخ است و لذا گزینه‌های انحرافی به‌عنوان گزینه‌های قابل‌قبول تنها زمانی مورد ارزیابی قرار می‌گیرد که آزمودنی قادر به حل سؤال نیست. به‌بیان‌دیگر، این رویکرد ضمن تفکیک دو منبع DIF، با بکار بردن یک استراتژی دومرحله‌ای روشن می‌سازد چرا یک سؤال ممکن است DIF داشته باشد اما DDF نشان ندهد (Suh & Talley, 2015).

بی‌شک، مشخص نمودن گزینه‌های انحرافی به‌عنوان علل DIF می‌تواند اطلاعات سودمندی برای تعدیل یا جایگزینی و نیز طراحی سؤال‌های جدید فراهم کند. هرچند تاکنون چندین روش برای شناسایی DDF در ادبیات روان‌سنجی پیشنهادشده است اما هنوز به‌طورقطع، مزیت این روش‌ها و کارایی نسبی آن‌ها تحت شرایط گوناگون معین نشده است و تحقیقات محدود انجام‌شده نیز غالباً نتایج ناهمسانی را گزارش کرده‌اند (Suh & Talley, 2015). هرچند به نظر می‌رسد رویکرد لوجیت آشیانه‌ای مبتنی بر چارچوب «تقسیم‌بر گزینه‌های انحرافی»، مزایای متعددی نسبت به سایر رویکردها در تعیین علل احتمالی DIF فراهم کند (Suh & Bolt, 2011)، اما هنوز شواهد کافی برای این ادعا وجود ندارد؛ لذا پژوهش حاضر به دنبال آن بود تا با استفاده از داده‌های شبیه‌سازی‌شده که به مطالعات مونت‌کارلو[1] (MC) معروف‌اند و نیز داده‌های واقعی، به بررسی کارایی مدل دو پارامتری لوجیت آشیانه‌ای (2PL-NLM) در تشخیص کنش افتراقی گزینه‌های انحرافی در شرایط مختلف و رابطه آن با کارکرد افتراقی سؤال بپردازد. دلیل استفاده از داده‌های شبیه‌سازی‌شده این بود که در بسیاری از موقعیت‌ها دستیابی به داده‌های واقعی تحت شرایط کنترل‌شده و

1. Monte Carlo Studies

تکرارهای فراوان، عملاً غیرممکن است؛ درحالی‌که مطالعات MC ضمن مدیریت داده‌ها، امکان کنترل یا دست‌کاری متغیرهایی مانند توزیع توانایی، ویژگی‌های سؤال و مطالعه اثرات چندین عامل در یک‌زمان را فراهم می‌کند (Harwell et al., 1996). در این مطالعه، هدف از کاربرد داده‌های شبیه‌سازی‌شده، ارزیابی خطای نوع اول و توان آزمون و هدف استفاده از داده‌های واقعی، بررسی رابطه کارکرد افتراقی سؤال و کارکرد گزینه‌های انحرافی بود. برای نیل به این اهداف، سؤالات زیر مدنظر قرار گرفت:

-آیا شناسایی کارکرد افتراقی سؤال با استفاده از مدل دو پارامتری لوجیت آشیانه‌ای تحت تأثیر توزیع توانایی گروه‌ها (یکسان و متفاوت)، ویژگی‌های سؤال (دشواری و تشخیص)، اندازه‌ی DDF (متوسط، بزرگ) و اندازه‌ی DIF (کوچک، متوسط، بزرگ) قرار می‌گیرد؟

-چه رابطه‌ای بین کارکرد افتراقی سؤال و کارکرد افتراقی گزینه‌های انحرافی در آزمون‌های واقعی وجود دارد؟

روش

برای ارزیابی عملکرد گزینه‌های انحرافی به‌عنوان منابع احتمالی DIF یک مطالعه تجربی (آزمایشی) با استفاده از داده‌های شبیه‌سازی‌شده و به دنبال آن، یک تحلیل ثانویه با استفاده از داده‌های واقعی انجام گرفت؛ بنابراین روش پژوهش ترکیبی از روش‌های مداخله‌ای (آزمایشی) و توصیفی-تحلیلی بود.

جامعه آماری و گروه نمونه: جامعه آماری برای مطالعه شبیه‌سازی‌شده عبارت بود از کلیه موقعیت‌ها و شرایطی که در آن از سؤال‌های چهارگزینه‌ای در آزمون‌های سرنوشت‌ساز استفاده می‌شود. حجم نمونه برای مطالعه شبیه‌سازی ۲۰۰۰ نفر (۱۰۰۰ گروه کانونی و ۱۰۰۰ گروه مرجع) در نظر گرفته شد. برای تحلیل ثانویه با داده‌های واقعی، جامعه آماری شامل کلیه دانش‌آموزانی بود که در کنکور سال ۹۷ در رشته تجربی شرکت نموده و حداقل به یک سؤال آزمون ریاضی پاسخ داده بودند. از این جامعه نیز، نمونه‌ای شامل ۱۰۰۰ مرد (گروه مرجع) و ۱۰۰۰ زن (گروه کانونی) که هریک به‌طور تصادفی از داده‌های ارائه‌شده توسط سازمان سنجش کشور استخراج‌شده بود، استفاده شد. دلیل استفاده از حجم نمونه‌های یکسان برای این پژوهش ازآنجا ناشی می‌شود که مطالعات تجربی (مثلاً، کاپلان و جورج،۱۹۹۵) و نتایج مطالعات شبیه‌سازی‌شده (مثلاً، گونزالس و همکاران، ۲۰۰۶) نشان داده‌اند که استفاده

از حجم نمونه‌های نابرابر گروه‌ها، توان آماری شاخص‌های ردگیری DIF را کاهش می‌دهد (Feinberg & Rubright, 2016).

ابزار اندازه‌گیری: ابزار جمع‌آوری داده‌های شبیه‌سازی‌شده یک آزمون ساختگی ۳۶ سؤالی چهارگزینه‌ای (متشکل از ۶ سؤال موردمطالعه و ۳۰ سؤال لنگر) بود که تحت ۵۰ تکرار بر اساس مدل دو پارامتری لوجیت آشیانه‌ای و با استفاده از بسته نرم‌افزاری mcIRT در محیط R (Reif, 2015) ایجاد گردید. برای جمع‌آوری داده‌های تجربی از فرم D خرده آزمون «ریاضی» از مجموعه آزمون‌های تخصصی زیرگروه علوم تجربی کنکور سرا سری سال ۹۷ استفاده شد. این آزمون دارای ۳۰ سؤال چهارگزینه‌ای بود؛ اما به علت حجم بالای داده‌های گم‌شده[1] در نمونه‌های استخراجی (به‌طور متوسط ۵۸ درصد داده‌های هر سؤال در گروه مردان و به همین میزان در گروه زنان فاقد پاسخ بود) که امکان کاربرد مدل‌های مبتنی بر IRT را منتفی می‌ساخت، داده‌های بدون پاسخ به‌عنوان یک طبقه پاسخ جدید وارد تحلیل شد به‌طوری‌که آزمون همانند یک آزمون پنج گزینه‌ای (یک گزینه صحیح و چهار گزینه غلط) مورد تحلیل قرار گرفت. علت انتخاب این خرده آزمون ازآنجا ناشی می‌شد که بر طبق گزارش‌های ارائه‌شده، در کنکور سال ۹۷، گروه آزمایشی علوم تجربی بیشترین (۶۰ درصد) متقاضیان را به خود اختصاص داده بود؛ و این گروه از پایین‌ترین نرخ پذیرش (۳۰ درصد) و درنتیجه از بالاترین میزان رقابت بین داوطلبان، نسبت به سایر گروه‌ها برخوردار بود. یکی از مهم‌ترین مواد امتحانی این گروه آزمایشی، خرده آزمون ریاضی است که در همه زیرگروه‌های آزمایشی علوم تجربی مورد سؤال واقع می‌شود و از ضریب بالایی برخوردار است. لذا موفقیت در این خرده آزمون، تأثیر زیادی بر موفقیت کلی داوطلبان دارد. از طرف دیگر استراتژی غالب در پاسخگویی به سؤال‌های آزمون ریاضی، منطبق بر استراتژی «حل مسئله» است (Hutchinson, 1991) که با رویکرد مفهومی DDF (تقسیم‌بر گزینه‌های انحرافی) همسو است.

شرایط ایجاد داده‌های شبیه‌سازی‌شده

متغیر وابسته: در این مطالعه نرخ‌های تشخیصِ صحیح و غلطِ کارکرد افتراقی، به‌عنوان متغیر اصلی برونداد در نظر گرفته شد. به عبارت دقیق‌تر پیامدهای موردمطالعه عبارت بودند از:۱) نرخ خطای نوع اول برای تعیین وجود کنش افتراقی (DDF/DIF) هنگامی‌که

1. missing value

DDF/DIF ی وجود ندارد و ۲) توان ردگیری DDF/DIF زمانی که DDF/DIF وجود دارد. متغیرهای مستقل یا عواملی که انتخاب یا مورد دست‌کاری قرار گرفت، عبارت بودند از: پارامترهای سؤال، توزیع توانایی، وجود یا نبود کارکرد افتراقی، بزرگی DIF و بزرگی DDF، همچنین در این مطالعه متغیرهایی چون حجم نمونه و طول آزمون برای دو گروه مرجع و کانونی یکسان انتخاب شد.

توزیع توانایی: تفاوت‌های توزیع توانایی می‌تواند بر ردگیری DIF اثر بگذارد (Judin & Girel, 2001)؛ بنابراین به‌منظور شبیه‌سازی موقعیتی که در آن سطح توانایی گروه‌ها یکسان باشد (نبود اثر[1]) از توزیع استاندارد $\theta \sim Normal(0, 1)$ برای هر دو گروه و برای شبیه‌سازی موقعیتی که در آن، یک گروه توانایی بیشتری نسبت به گروه دیگر داشته باشد (وجود اثر)، از توزیع $\theta \sim Normal(0, 1)$ برای گروه مرجع و از توزیع $\theta \sim Normal(-.5, 1)$ برای گروه کانونی استفاده شد.

ویژگی سؤال‌های موردمطالعه: داده‌های مربوط به ۶ سؤال موردمطالعه، بر اساس سه شرط: (۱) داده‌هایی بدون DIF و بدون DDF، (۲) داده‌های با DIF تنها و (۳) داده‌هایی با هردوی DIF و DDF ایجاد شد. برای داده‌های بدون DIF، دو مقدار پارامتر شیب ۱/۲۵، α = ۰/۷۵ با سه اندازه پارامتر دشواری ۱/۵، ۰، ۱/۵- = β برای طبقه پاسخ صحیح ترکیب شد و پارامترهای شش سؤال حاصل، برای هر دو گروه کانونی و مرجع به‌طور یکسان تنظیم گردید. مقادیر پارامترهای گزینه‌های انحرافی این ۶ سؤال، با توجه به محدودیت $\sum_{v=1}^{m} l_{iv} = 0$ و $\sum_{v=1}^{m} z_{iv} = 0$ به میزان ۰/۲۶، ۰/۲۸-، ۰/۰۲ = l_v برای شیب و به میزان ۰/۲۲-، ۰/۱۴-، ۰/۳۶ = z_v برای عرض از مبدأ اختیار شد و برای همه شش سؤال موردمطالعه در هردو گروه مرجع و کانونی ثابت در نظر گرفته شد. مقادیر این پارامترها از مطالعه Suh and Bolt (2011) به‌دست‌آمده و دلیل انتخاب این مقادیر آن بود تا معرف سطوح مختلف پارامترهایی باشند که در موقعیت‌های عملی آزمودن مرسوم‌اند. بر این اساس، شش سؤال بدون DIF (و بدونDDF) از ترکیب سه پارامتر دشواری [بالا (H)، متوسط (M)، کوچک (L)] و دو پارامتر شیب (بالا (H)، اندک (L)) به دست آمد. برای سؤال‌های DIF دار: سه سطح DIF ناچیز (Δa = ۰.۲۵)، متوسط (Δa = ۰.۵) و بزرگ (Δa = ۱) و برای داده‌های DDF دار: دو سطح DDF متوسط ($\Delta\zeta$ = ۰/۴) و بزرگ ($\Delta\zeta$ = ۱/۲) شبیه‌سازی شد.

1. no impact

همچنین DIF در پارامتر شیب سؤال به میزان Δb = ۰/۳ و DDF در پارامتر شیب گزینه‌های انحرافی نیز به میزان $\Delta\lambda$ = ۰/۳ به‌طور هم‌زمان با DIF در پارامتر دشواری و عرض از مبدأ معرفی گردید چراکه این شکل از کنش افتراقی (غیریکنواخت) به‌کرات به‌ویژه هنگامی‌که DDF نیز حضور دارد، رخ می‌دهد (Penfield, 2010). درمجموع از ترکیب شرایط مختلف ۶ سؤال بدون DIF (و بدون DDF)، ۱۸ سؤال DIF دار و ۳۶ سؤال DIF+DDF دار متفاوت شبیه‌سازی شد.

سؤال‌های لنگر: رویکردهای ردگیری DIF و DDF نیازمند تفکیک اساسی بین دو زیرمجموعه از سؤال‌ها، یعنی یک زیرمجموعه جور شده (مجموعه سؤال‌های لنگر[1]) برای مشخص کردن مقیاس مشترک اندازه توانایی بین گروه‌ها و یک زیرمجموعه مورد ظن برای بررسی کنش افتراقی (سؤال‌های موردمطالعه) است (امبرتسون و رایس، ۲۰۰۰، ترجمه شریفی، ۱۳۸۸). در این مطالعه برای ایجاد سؤال‌های لنگر، ابتدا ۵۰ سؤال چهارگزینه‌ای مبتنی بر 2PL-NLM و با استفاده از توزیع‌های: پارامترهای شیب با تابع یکنواخت؛ $\alpha \sim \text{unif}$ (.75, 2) و دشواری با تابع یکنواخت؛$\beta \sim \text{unif}\ (-2.5, 2.5)$ برای طبقه پاسخ صحیح؛ و پارامترهای شیب با تابع یکنواخت؛ $\lambda \sim \text{unif}\ (-2, 2)$ و عرض از مبدأ با تابع یکنواخت $\xi \sim \text{unif}\ (-2, 2)$ برای طبقات گزینه‌های انحرافی و با اعمال محدودیت $\sum_{v=1}^{m} z_{iv} = 0$ و $\sum_{v=1}^{m} l_{iv} = 0$ برای آن طبقات، شبیه‌سازی شد. سپس با استفاده از روش‌های پالایش ازسرگیرانه[2] (Kim & Cohen, 1995) به کمک آزمون LR درنهایت ۳۰ سؤال که فاقد DIF و DDF بودند به‌عنوان لنگر نهایی تعیین شد.

تعداد تکرار[3]: تعداد نسخه‌ها در مطالعات MC مترادف با حجم نمونه در تحقیقات تجربی است. هنگامی‌که هدف، مقایسه روش‌شناختی‌های مبتنی بر IRT است (مثلاً مقایسه تعداد سؤال‌های DIF داری که به‌درستی توسط روش‌ها، ردگیری‌شده است) توزیع‌های نمونه‌گیری تجربی لزوماً معمول نیست و تعداد نسخه‌های اندک (مثلاً ۱۰ مورد) ممکن است کافی باشد (Feinberg & Rubright, 2016)، همچنین بر اساس مطالعات استون (۱۹۹۳) و پیشنهاد Harwell و همکاران (1996)، حداقل ۲۵ نسخه برای داشتن یک توان مناسب

1. anchor item set
2. iterative purification procedures
3. Number of Replications

برای ردگیری اثرات ضروری است. (همان) لذا در این مطالعه، برای هر یک از ترکیب‌های مطالعه MC، از ۵۰ تکرار استفاده شد.

روش تحلیل داده‌ها در ادامه بیان شده است:

لوجیت دو پارامتری آشیانه‌ای[1]: Suh & Bolt (2010) مبتنی بر چارچوب «تقسیم‌بر گزینه‌های انحرافی» روشی را با استفاده از مدل لوجیت آشیانه‌ای به‌منظور مدل‌سازی پاسخ‌ها در سؤال‌های چندگزینه‌ای مطرح کردند و آن را برای IRT تطبیق دادند به‌طوری‌که امکان بازرسی عملکرد گزینه‌های انحرافی را مستقل از DIF فراهم می‌سازد. این روش شامل دو مؤلفه مجزای: (۱) مشخص کردن احتمال پاسخ صحیح و (۲) مشخص کردن احتمال انتخاب گزینه انحرافی مشروط بر پاسخ غلط است. مطابق Suh and Bolt (2011) تحت این مدل، احتمال آنکه آزمودنی با توانایی θ_j گزینه صحیح را در سؤال i انتخاب کند می‌تواند به شکل مدل لجستیک دو پارامتری، مطابق زیر، الگو پردازی شود.

معادله ۱

$$P_i(\theta_j) = \frac{1}{1 + \exp^{-a_i(\theta_j - b_i)}}$$

که در آن b_i نشانگر پارامتر دشواری و a_i پارامتر شیب برای سؤال i است.

و احتمال آنکه آزمودنی، هر طبقه پاسخ انحرافی v ($v = 1, 2, .., m$) را انتخاب کند، با استفاده از الگو پردازی مدل پاسخ اسمی Bock (1972) به‌صورت حاصل‌ضرب احتمال یک پاسخ غلط در احتمال انتخاب یک طبقه پاسخ انحرافیِ v مشروط بر پاسخ غلط به دست می‌آید (معادله ۲).

معادله ۲

$$P_{iv}(\theta_j) = \left[\frac{1}{1 + \sum_{v=1}^{m} \exp^{z_{iv} + l_{iv}\theta_j}}\right]\left[\exp^{z_{iv} + l_{iv}\theta_j}\right]$$

که در آن ζ و λ به ترتیب عرض از مبدأ و شیب طبقه پاسخ انحرافی است. پرانتز دوم در معادله ۲ همان مدل پاسخ اسمی Bock (1972) است؛ با این تفاوت که مخرج کسر تنها بر اساس طبقات پاسخ انحرافی تعریف‌شده است (Suh & Talley, 2015) مطابق رویکرد Bock (1972)، برای اطمینان از شناسایی و مشخص بودن مدل، برای طبقات پاسخ انحرافی محدودیت $\sum_{v=1}^{m}(\zeta_{iv} + \lambda_{iv}\theta_j) = 0$ اعمال می‌شود (Desjardins & Bulut, 2018) که

1. 2PL-nested logit model (2PL-NLM)

به معنی آن است که مجموع پارامترهای دشواری و مجموع پارامترهای تشخیص برای گزینه‌های انحرافی برابر صفر در نظر گرفته می‌شود. یک بسط دوگروهی از مدل آشیانی دو پارامتری، امکان اینکه هردو پارامترهای طبقه پاسخ صحیح و پارامترهای طبقه انحرافی در بین گروه‌ها متفاوت باشد را فراهم می‌سازد. در این شرایط، معادله (۱) می‌تواند به‌صورت زیر بازنویسی شود (Suh & Bolt, 2011):

معادله ۳

$$p(u_{ij} = 1 | q_j, G = g) = \frac{\exp(b_{ig} + a_{ig} q_j)}{1 + \exp(b_{ig} + a_{ig} q_j)}$$

که در آن G شاخص گروه است. به همین ترتیب احتمال انتخاب گزینه انحرافی (معادله ۲) به‌صورت زیر بیان می‌شود:

معادله ۴

$$P(d_{ijv} = 1 | u_{ij} = 0, q_j, G = g) = \frac{\exp(Z_{igv}(q_j))}{\sum_{k=1}^{m} \exp(Z_{igv}(q_j))}$$

که در آن $Z_{ij}(q_j) = z_{iv} + l_{iv}(q_j)$ است. مطابق Suh and Bolt (2011)، جهت ارزیابی تفاوت پارامترهای مربوط به منحنی پاسخ در بین گروه‌ها، می‌توان توابع درستنمایی (LR) مدل‌های مختلف را مورد مقایسه قرارداد. برای این منظور، سه مدل سلسله مراتبی با محدودیت‌های[1] مختلف مورد ملاحظه قرار می‌گیرد: (۱) یک مدل فشرده[2] که در آن‌همه پارامترهای سؤال موردمطالعه در بین گروه‌ها یکسان برآورد می‌شود. (۲) اولین مدل افزوده[3] که در آن تنها پارامترهای طبقه پاسخ انحرافیِ سؤالِ موردمطالعه در بین گروه‌ها یکسان در نظر گرفته می‌شود و (۳) دومین مدل افزوده که در آن هیچ‌یک از پارامترهای سؤالِ موردمطالعه در بین گروه‌ها یکسان در نظر گرفته نمی‌شود و به‌صورت آزاد برآورد می‌شوند. با محاسبه دو برابر تفاوت منفی اندازه لگاریتم درستنمایی $G^2 = -2 \log L_1 - (-2 \log L_2)$ بین مدل فشرده با اولین مدل افزوده (آزمون ۱) می‌توان وجود DIF را آزمود. همچنین، با محاسبه آماره G^2 ی که اولین و دومین مدل افزوده را مقایسه می‌کند (آزمون ۲) می‌توان حضور DDF را ارزیابی نمود. به‌طور خاص، اگر برای سؤال موردمطالعه هردو آزمون ۱ و

1. constraints
2. compact
3. augmented

۲ معنی‌دار شود، نشانگر وجود هردوی DIF و DDF در سؤال است و گزینه‌های انحرافی احتمالاً به‌عنوان بخشی از علت DIF عمل می‌کنند (Suh & Bolt, 2011).

ارائه اندازه اثر[1] که شاخصی از بزرگی اثر یافت شده است، در آزمون فرض‌های آماری، لازم است چراکه حجم نمونه‌های کوچک می‌تواند اثرات آماری موردنظر را پنهان کند درحالی‌که نمونه‌های با حجم بزرگ می‌تواند سبب معنی‌داری اثراتی کاملاً کوچک و بی‌معنی شود (کیرک، ۱۹۹۶). برای تعیین اندازه اثر کنش افتراقی در این مطالعه از شاخص معیار یابی[2] (شاخص استانداردSTD) Dorans و همکاران (1992) استفاده شد.

یافته‌ها

داده‌های شبیه‌سازی‌شده در ادامه بررسی شده است:

با استفاده از آماره G^2 که دارای توزیع کای اسکور با درجه آزادی برابر با تفاوت در تعداد پارامترهای برآورد شده بین دو مدل است، یک استراتژی دومرحله‌ای (آزمون ۱ و آزمون ۲) برای مشخص نمودن سؤال‌های دارایDIF و DDF اجرا شد. برای آزمون ۱ یعنی آزمون وجود DIF، درجه آزادی برابر ۲ و برای آزمون ۲ یعنی آزمون وجودDDF درجه آزادی ۴ بود. لذا با توجه به سطح معنی‌داری ۰.۰۵، ارزش‌های بحرانی ۵/۹۹ و ۱۲/۵۹ به ترتیب برای آزمون‌های ۱ و ۲ ملاک عمل قرار گرفت. برای ترکیب آزمون‌های ۱ و ۲ با فرض استقلال آن‌ها، سطح معنی‌داری α =۰/۰۰۲۵ انتخاب شد (Suh & Bolt, 2001). ازآنجاکه ۵۰ نسخه تکرار صورت گرفته بود، تعداد مورد انتظار G^2 معنی‌دار به‌واسطه شانس در سطح معنی‌داری ۰/۰۵ و ۰/۰۰۲۵ به ترتیب برای هر سؤال موردمطالعه ۲ و ۰ (گرد شده) بود.

۱) داده‌های بدون DIF و توزیع یکسان دو گروه

جدول ۱، تعداد و درصد G^2 معنی‌دار برای هریک از ۶ سؤال موردمطالعه تحت توزیع یکسان توانایی را نشان می‌دهد. بر اساس آزمون ۱، نرخ خطا برای دو سؤال ۳ و ۵ (LH و HM) بیشتر از ارزش مورد انتظار و برای سؤال‌های دیگر تا حدودی کمتر از ارزش مورد انتظار بود. بااین‌حال، متوسط نرخ خطا کمتر از ارزش مورد انتظار (۰/۰۵) به دست آمد. در مورد آزمون ۲، یعنی آزمون DDF، برای دو سؤال موردمطالعه (LM، LH) تا حدودی تورم نرخ خطا دیده می‌شود؛ اما همچنان متوسط نرخ خطا، به میزان جزئی کمتر از ارزش مورد انتظار

1. effect size
2. Standardization Index

است. کارکرد ترکیبی دو آزمون، یعنی آزمون هم‌زمان DDF /DIF نیز نشان داد که نرخ‌های خطا برای همه سؤال‌های موردمطالعه، برابر با ارزش مورد انتظار است

جدول ۱. تعداد و (درصد) G^2 معنی‌دار برای ٦ سؤال موردمطالعه تحت داده‌های بدون DIF و توزیع یکسان

سؤال	نوع سؤال	پارامترهای طبقه پاسخ β	پارامترهای طبقه پاسخ α	آزمون ۱ (DIF)	آزمون ۲ (DDF)	آزمون ۱ و ۲ (DIF + DDF)
۱	LL	−۱/۵	۰/۷۵	۲ (٤)	۲ (٤)	۰ (۰)
۲	LM	۰	۰/۷۵	۱ (۲)	۳ (٦)	۰ (۰)
۳	LH	۱/۵	۰/۷۵	۳ (٦)	۳ (٦)	۰ (۰)
٤	HL	−۱/۵	۱/۲۵	۲ (٤)	۲ (٤)	۰ (۰)
٥	HM	۰	۱/۲۵	۳ (٦)	۲ (٤)	۰ (۰)
٦	HH	۱/۵	۱/۲۵	۲ (٤)	۲ (٤)	۰ (۰)

درصدها از طریق تقسیم تعداد G^2 معنی‌دار بر تعداد تکرارها (۵۰) به دست می‌آید

۲) داده‌های بدون DIF و توزیع متفاوت دو گروه

جدول ۲ تعداد و درصد G^2 معنی‌دار تحت شرایط توزیع متفاوت توانایی را نشان می‌دهد. بر اساس محتویات این جدول، تعداد آماره‌های معنی‌دار نسبت به جدول ۱ افزایش‌یافته است به‌طوری‌که متوسط نرخ خطا برای آزمون DIF و DDF به میزان جزئی بیشتر از ارزش مورد انتظار بود. برای آزمون هم‌زمان DDF /DIF نیز نرخ خطا برای یک سؤال بیشتر از رزش مورد انتظار بود.

جدول ۲. تعداد و (درصد) G^2 معنی‌دار برای ٦ سؤال موردمطالعه تحت داده‌های بدون DIF و توزیع متفاوت

سؤال	نوع سؤال	پارامترهای طبقه پاسخ β	پارامترهای طبقه پاسخ α	آزمون ۱ (DIF)	آزمون ۲ (DDF)	آزمون ۱ و ۲ (DIF + DDF)
۱	LL	−۱/۵	۰/۷۵	۵ (۱۰)	۲(٤)	۱ (۲)
۲	LM	۰	۰/۷۵	۲ (٤)	۲ (٤)	۰ (۰)
۳	LH	۱/۵	۰/۷۵	۱ (۲)	٤ (۸)	۰ (۰)
٤	HL	−۱/۵	۱/۲۵	۲ (٤)	۲ (٤)	۰ (۰)
٥	HM	۰	۱/۲۵	۵ (۱۰)	۲ (٤)	۰ (۰)
٦	HH	۱/۵	۱/۲۵	۵ (۱۰)	٦ (۱۲)	۰ (۰)

درصدها از طریق تقسیم تعداد G^2 معنی‌دار بر تعداد تکرارها (۵۰) به دست می‌آید

۳) داده‌های DIF دار و توزیع یکسان دو گروه

نرخ‌های رد برای مجموعه داده‌های DIF دار تحت شرایط توزیع یکسان در جدول ۳ منعکس‌شده است. در این جدول، ستون‌های اول تا سوم برای آزمون ۱، نرخ‌های رد آماره G^2 برای آزمون DIF (رد فرضِ صفرِ نبود کنش افتراقی) در سطح $\alpha = ۰/۰۵$ را نشان می‌دهد. همان‌طور که انتظار می‌رفت، با افزایش سطح DIF در پارامتر دشواری از مقدار ۰/۲۵، ۰/۵ تا ۱/۰ (DIF-L→DIF-M →DIF-H)، نرخ‌های رد فرض صفر نیز افزایش یافت. به‌طوری‌که تحت شرایط DIF-L، نرخ‌های رد دامنه‌ای از ۰/۱۲ تا ۰/۸۴ و تحت شرایط DIF-M، نرخ‌های رد دامنه‌ای از ۰/۶۴ تا ۰/۹۸ نشان داد، درحالی‌که تحت شرایط DIF-H، آزمون به‌درستی همه سؤال‌های DIF دار را تشخیص داد. از طرف دیگر، تحت شرایط DIF-L و تا حدودی تحت شرایط DIF-M در آزمون ۱، دو الگوی تفسیر را می‌توان مشاهده کرد. یک الگو آن است که به‌طور متوسط، نرخ رد تحت شرایط شیب اندک (α= ۰/۷۵) به مقدار جزئی بیشتر از شرایط شیب بالا (α= ۱/۲۵) ظاهر می‌شود؛ و الگوی دیگر آنکه با افزایش پارامتر دشواری سؤال (β) نرخ رد فرض صفر افزایش می‌یابد. سه ستون بعدی مربوط به آزمون ۲، منعکس‌کننده نرخ خطای نوع اول برای آزمون DDF است. همان‌طور که مشاهده می‌شود اغلب شرایط تا حدودی منجر به تورم خطا می‌شود.

جدول ۳. تعداد و (درصد) G^2 معنی‌دار برای ٦ سؤال موردمطالعه تحت داده‌های DIF دار و توزیع یکسان

سؤال	نوع سؤال	آزمون ۱ (DIF)			آزمون ۲ (DDF)		
		DIF-L	DIF-M	DIF-H	DIF-L	DIF-M	DIF-H
۱	LL	۸ (۱٦)	۳۷ (۷٤)	۵۰ (۱۰۰)	۳ (٦)	۷ (۱٤)	۱ (۲)
۲	LM	۲۲ (٤٤)	٤۸ (۹٦)	۵۰ (۱۰۰)	٤ (۸)	۱ (۲)	۱ (۲)
۳	LH	٤۲ (۸٤)	٤۹ (۹۸)	۵۰ (۱۰۰)	۲ (٤)	۳ (٦)	۳ (٦)
٤	HL	٦ (۱۲)	۳۲ (٦٤)	۵۰ (۱۰۰)	۲ (٤)	٤ (۸)	٤ (۸)
۵	HM	۱۸ (۳٦)	٤۹ (۹۸)	۵۰ (۱۰۰)	٤ (۸)	۳ (٦)	۳ (٦)
٦	HH	۳۷ (۷٤)	٤۹ (۹۸)	۵۰ (۱۰۰)	۲ (٤)	٤ (۸)	۲ (٤)

درصدها از طریق تقسیم تعداد G^2 معنی‌دار بر تعداد تکرارها (۵۰) به دست می‌آید

۴) داده‌های DIF دار و توزیع متفاوت دو گروه

برای مجموعه داده‌های DIF دار و تحت شرایط توزیع متفاوت توانایی نیز همانند جدول قبل، با افزایش سطح DIF، نرخ‌های رد فرض صفر افزایش می‌یابد؛ همچنین شرایط DIF-L و DIF-M مربوط به آزمون ۱، دو الگوی قبلی را مجدداً آشکار می‌کند؛ یعنی با افزایش پارامتر دشواری و کاهش شیب سؤال، نرخ رد افزایش می‌یابد. برای آزمون وجود DDF نیز، اغلب شرایط، به‌ویژه شرایط شیب بالا تا حدودی منجر به تورم خطای نوع اول می‌شود (جدول ۴).

جدول ٤. تعداد و (درصد) G^2 معنی‌دار برای ٦ سؤال موردمطالعه تحت داده‌های DIF دار و توزیع متفاوت

		آزمون ۱ (DIF)			آزمون ۲ (DDF)		
سؤال	نوع سؤال	DIF-L	DIF-M	DIF-H	DIF-L	DIF-M	DIF-H
۱	LL	۳۱ (٦۲)	٤٤ (۸۸)	۵۰ (۱۰۰)	۱ (۲)	۲ (٤)	۲ (٤)
۲	LM	٤٤ (۸۸)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۲ (٤)	۱ (۲)	۲ (٤)
۳	LH	٤٦ (۹۲)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۳ (٦)	۳ (٦)	٤ (۸)
٤	HL	۲۵ (۵۰)	۳۷ (۷٤)	۵۰ (۱۰۰)	۹ (۱۸)	٦ (۱۲)	۳ (٦)
۵	HM	٤٤ (۸۸)	٤۸ (۹٦)	۵۰ (۱۰۰)	۲ (٤)	۵ (۱۰)	۳ (٦)
٦	HH	٤٦ (۹٦)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۳ (٦)	۲ (٤)	۳ (٦)

درصدها از طریق تقسیم تعداد G^2 معنی‌دار بر تعداد تکرارها (۵۰) به دست می‌آید

به‌طورکلی مقایسه توزیع‌های یکسان و متفاوت نشان داد؛ هرچند نرخ‌های رد تحت شرایط توزیع متفاوت توانایی، همان الگوهای توزیع یکسان توانایی را نشان می‌دهد، اما نرخ‌های رد در شرایط توزیع متفاوت بیشتر از شرایط توزیع یکسان است.

۵) داده‌های DIF+DDF دار و توزیع یکسان دو گروه

جدول ۵ نرخ‌های رد فرض صفر برای مجموعه داده‌های دارای DIF+DDF تحت شرایط توزیع یکسان را نشان می‌دهد. در این جدول ستون‌های مربوط به آزمون ۱ (DIF) و آزمون ۲ (DDF) نرخ رد را در سطح α = ۰/۰۵ نشان می‌دهد، درحالی‌که دو ستون متعلق به آزمون ترکیبی آزمون ۱ و ۲ (DIF + DDF)، نرخ رد را در سطح α = ۰/۰۰۲۵ نشان می‌دهد. همان‌طور که مشاهده می‌شود، در اینجا نیز به‌طور متوسط با افزایش سطح DIF، یا با افزایش سطح DDF، نرخ رد-صرف نظر از جند مورد- افزایش می‌یابد. همچنین برای

آزمون ۱، دو الگوی مشاهده‌شده در جداول قبل مجدداً آشکار می‌شود؛ یعنی، به‌طور متوسط، نرخ رد تحت شرایط شیب اندک به مقدار جزئی بیشتر از شرایط شیب بالا ظاهر می‌شود و با افزایش پارامتر دشواری (β)، نرخ رد افزایش می‌یابد؛ اما برخلاف آزمون ۱، نرخ‌های رد برای آزمون ۲ با کاهش سطح عرض از مبدأ و یا با افزایش شیب سؤال، تمایل به افزایش را نشان می‌دهد (جدول ۵). چنین نتایجی قابل‌انتظار است، چراکه با افزایش دشواری و قدرت تمیز سؤال، پاسخ‌های غلط برای افراد با توانایی اندک جذابیت بیشتری پیدا می‌کند.

جدول ۵. تعداد و (درصد) G^2 معنی‌دار برای ٦ سؤال موردمطالعه تحت داده‌های DIF+DDF دار و توزیع یکسان

		آزمون ۱ (DIF)		آزمون ۲ (DDF)		آزمون ۱ و ۲ (DIF + DDF)	
کنش	سؤال	DDF-M	DDF-H	DDF-M	DDF-H	DDF-M	DDF-H
DIF-L	LL	۷ (١٤)	٣٣ (٦٦)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)
	LM	٢٣ (٤٦)	٤٩ (٩٨)	٤٩ (٩٨)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)
	LH	٣٩ (٧٨)	٤٩ (٩٨)	٤٠ (٨٠)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)
	HL	۵ (۱۰)	٢٩ (٥٨)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	٤٩ (٩٨)	۵۰ (۱۰۰)
	HM	١٧ (٣٤)	٤٥ (٩٠)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	٤٩ (٩٨)	۵۰ (۱۰۰)
	HH	٣٥ (٧٠)	٤٩ (٩٨)	٤٦ (٩٢)	۵۰ (۱۰۰)	٤٦ (٩٢)	۵۰ (۱۰۰)
DIF-M	LL	٣٦ (٨٢)	٤١ (٨٢)	٤٢ (٨٤)	۵۰ (۱۰۰)	٤٦ (٩٢)	۵۰ (۱۰۰)
	LM	٤٩ (٩٨)	٤٩ (٩٨)	٤٧ (٩٤)	۵۰ (۱۰۰)	٤٩ (٩٨)	۵۰ (۱۰۰)
	LH	۵۰ (۱۰۰)	۵۰ (۱۰۰)	٣٥ (٧٠)	۵۰ (۱۰۰)	٤٩ (٩٨)	۵۰ (۱۰۰)
	HL	٣٢ (٦٤)	٣٩ (٧٨)	٤٦ (٩٢)	۵۰ (۱۰۰)	٤٣ (٨٦)	۵۰ (۱۰۰)
	HM	٤٨ (٩٦)	٤٨ (٩٦)	٤٨ (٩٦)	۵۰ (۱۰۰)	٤٨ (٩٦)	۵۰ (۱۰۰)
	HH	۵۰ (۱۰۰)	۵۰ (۱۰۰)	٤٠ (٨٠)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)
DIF-H	LL	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)
	LM	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)
	LH	۵۰ (۱۰۰)	۵۰ (۱۰۰)	٤٢ (٨٤)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)
	HL	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)
	HM	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)
	HH	۵۰ (۱۰۰)	۵۰ (۱۰۰)	٤٦ (٩٢)	۵۰ (۱۰۰)	۵۰ (۱۰۰)	۵۰ (۱۰۰)

درصدها از طریق تقسیم تعداد G^2 معنی‌دار بر تعداد تکرارها (۵۰) به دست می‌آید

۶) داده‌های DIF+DDF دار و توزیع متفاوت دو گروه

جدول ۶ نرخ‌های رد برای داده‌های دارای DIF+DDF تحت شرایط توزیع متفاوت را نشان می‌دهد. در این شرایط نیز، به‌طور متوسط با افزایش سطح DIF، یا DDF نرخ رد افزایش می‌یابد. همچنین برای آزمون ۱ و آزمون ۲ الگوهای مشاهده‌شده در جدول ۵، مجدداً آشکار می‌شود؛ یعنی برای آزمون ۱، به‌طور متوسط، با دشوارتر شدن سؤال و یا کاهش پارامتر شیب، نرخ رد به مقدار جزئی افزایش می‌یابد و در آزمون ۲ برعکس با کاهش دشواری و افزایش شیب سؤال، نرخ‌های رد مقداری افزایش را نشان می‌دهد.

جدول ٦. تعداد و (درصد) G^2 معنی‌دار برای ٦ سؤال موردمطالعه تحت داده‌های DIF+DDF دار و توزیع متفاوت

		آزمون ۱ (DIF)		آزمون ۲ (DDF)		آزمون ۱ و ۲ (DIF + DDF)	
کنش	سؤال	DDF-M	DDF-H	DDF-M	DDF-H	DDF-M	DDF-H
DIF-L	LL	٣٤ (٦٨)	٣٦ (٧٢)	٤٨ (٩٦)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)
	LM	٤٩ (٩٨)	٤٨ (٩٦)	٤٨ (٩٦)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)
	LH	٤٧ (٩٤)	٤٨ (٩٦)	٤٢ (٨٤)	٥٠ (١٠٠)	٤٧ (٩٤)	٥٠ (١٠٠)
	HL	٣٠ (٦٠)	٣٢ (٦٤)	٤٩ (٩٦)	٥٠ (١٠٠)	٤٨ (٩٦)	٥٠ (١٠٠)
	HM	٤٣ (٨٦)	٤٢ (٨٤)	٥٠ (١٠٠)	٥٠ (١٠٠)	٤٩ (٩٨)	٥٠ (١٠٠)
	HH	٥٠ (١٠٠)	٥٠ (١٠٠)	٤٩ (٩٦)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)
DIF-M	LL	٤٦ (٩٢)	٤٩ (٩٨)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)
	LM	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)
	LH	٥٠ (١٠٠)	٥٠ (١٠٠)	٤٦ (٩٢)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)
	HL	٤٥ (٩٠)	٤٧ (٩٤)	٥٠ (١٠٠)	٤٩ (٩٨)	٥٠ (١٠٠)	٥٠ (١٠٠)
	HM	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)
	HH	٥٠ (١٠٠)	٥٠ (١٠٠)	٤٧ (٩٤)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)
DIF-H	LL	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)
	LM	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)
	LH	٥٠ (١٠٠)	٥٠ (١٠٠)	٤٤ (٨٨)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)
	HL	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)
	HM	٥٠ (١٠٠)	٥٠ (١٠٠)	٤٩ (٩٨)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)
	HH	٥٠ (١٠٠)	٥٠ (١٠٠)	٤١ (٨٢)	٥٠ (١٠٠)	٥٠ (١٠٠)	٥٠ (١٠٠)

درصدها از طریق تقسیم تعداد G^2 معنی‌دار بر تعداد تکرارها (۵۰) به دست می‌آید

داده‌های واقعی در ادامه بررسی شده است:

شاخص‌های توصیفی: میانگین نمرات ریاضی گروه نمونه مردان کمی بیش از ۱ نمره بالاتر از میانگین نمرات زنان گروه نمونه بود به‌طوری‌که؛ میانگین نمرات ریاضی برای مردان ۱۱/۸۴، با انحراف استاندارد ۵/۴۸ و ضریب پایایی ۰.۸۱؛ و میانگین نمرات ریاضی برای زنان ۱۰/۶۳ با انحراف استاندارد ۴/۷۱ و ضریب پایایی ۰.۸۶ به دست آمد. درمجموع، میانگین نمره آزمون ریاضی ۱۱/۲۴ و با انحراف استاندارد ۵/۱۵ و ضریب پایایی آزمون ۰/۸۴ به دست آمد.

بعدیت آزمون: در این مطالعه به‌منظور بررسی فرض تک‌بعدی بودن آزمون از روش‌های مختلفی چون نمودار اسکری، نسبت ارزش ویژه عامل اول، تحلیل عاملی غیرخطی (NOHARM) و تحلیل عاملی سؤال با اطلاعات کامل[1] (IFA) استفاده شد. یافته‌ها نشان داد که می‌توان شواهد تائید کننده‌ای مبنی بر اینکه داده‌های آزمون ریاضی از یک عامل کلی اشباع است و لذا کاربردIRT تک‌بعدی قابل توجیه و نتایج حاصل از آن معتبر است، ارائه داد.

نتایج آزمون نسبت درستنمایی در ادامه بیان شده است:

به‌منظور تعیین زیرمجموعه سؤال‌های لنگر، روش ازسرگیری (Lord, 1980) با استفاده از آزمون نسبت درستنمایی و رویکرد اختصاصی پیشنهادشده توسط Kim and Cohen(1995) به کار گرفته شد؛ و سؤال‌هایی به‌عنوان سؤال‌های لنگر انتخاب شد که هم ازنظر گزینه پاسخ صحیح و هم ازنظر گزینه‌های انحرافی فاقد کنش افتراقی باشند (آزمون ۱ و آزمون ۲). بر اساس نتایج نهایی حاصل از روش پالایش ازسرگیری (تکرارشونده)، ۱۲ سؤال به دلیلDIF و یک سؤال به دلیل DDF از مجموعه سؤال‌ها خارج شد و ۱۷ سؤال از ۳۰ سؤال آزمون فاقد DIF و یا DDF شناسایی شد. این سؤال‌ها به‌عنوان مجموعه سؤال‌های لنگر به کار گرفته شد تا ۱۳ سؤال باقیمانده بتواند از حیث DIF و DDF بر اساس مدل 2PL-MNL مورد آزمون قرار گیرد. جدول ۷ نتایج این آزمون‌ها (آزمون ۱ و آزمون ۲) را نشان می‌دهد.

1. Full-information item factor analysis

جدول ۷. آزمون نسبت درستنمایی مدل‌های 2PL و NRM برای سؤال‌های تحت مطالعه آزمون ریاضی

سؤال	مدل لوجیت آشیانه‌ای						مدل پاسخ اسمی	
	آزمون ۱ (DIF)		آزمون ۲ (DDF)		آزمون ۱ و ۲			
	*G*2	Sig.	*G*2	Sig.	Sig.		*G*2	Sig.
٤	٦/٦٦		١/٦٤				١٠/٣١	
٩	١١/٣٠		١٣/٥٣				٣٥/٦٠	
١٠	١٣/٤١		٨/٨٤				٢٦/٥٥	
١١	١٢/١٩		٦/٨٩				١٨/٩٦	
١٢	٠/١٤		٢٢/١٩				٢٧/٤٦	
١٣	٣٣/٢٤		١٩/٧٤				٥٩/٠٧	
١٦	١٥/١٣		٣/٩٩				٢٠/٨٧	
١٩	٢٤/٥٧		١٦/٨٢				٤٥/٨٥	
٢٠	٢٢/١١		١٣/١٧				٣٤/٢٦	
٢١	٠/١٦		١٩/٤٥				١٧/٦٣	
٢٣	٦/١٨		٧/١٩				١١/٤٢	
٢٤	١٢/٠٣		٨/٩٣				٢١/٣٥	
٢٦	١٢/٦٠		١٣/٢٠				٣٠/٢٢	

ازآنجاکه سؤال‌ها بر اساس یک نمونه بزرگ مورد تحلیل قرار گرفت، برخی سؤال‌ها ممکن بود برحسب شانس، معنی‌داری آماری نشان دهند، بنابراین سطح معنی‌داری ۰.۰۱ = α (ارزش بحرانی ۹/۲۱ و ۱۶/۸۱ به ترتیب برای آزمون ۱ و ۲) ملاک عمل قرار گرفت. همان‌طور که جدول ۷ نشان می‌دهد، بر اساس آزمون ۱، ۹ سؤال از ۱۳ سؤال موردمطالعه واجد DIF بود درحالی‌که بر اساس آزمون ۲ تنها ۴ سؤال به‌عنوان DDF دار شناسایی شد. همچنین بر اساس نتایج هر دو آزمون، ۲ سؤال (سؤال‌های ۱۳ و ۱۹) هردوی DIF و DDF را نشان داد.

در کنار نتایج حاصل از کاربرد مدل دو پارامتری لوجیت آشیانی، نتایج حاصل از آزمون نسبت درستنمایی تحت مدل پاسخ اسمی Bock (NRM) (1972) نیز با استفاده از همان سؤال‌های لنگر ارائه شد تا پایه‌ای برای مقایسه فراهم شود. برای این تحلیل، همه طبقات پاسخ و یک طبقه بدون پاسخ –داده‌های گم‌شده– هر سؤال به‌طور هم‌زمان برای کارکرد افتراقی مورد آزمون قرار گرفت؛ و نتایج آماره آزمون، با مقدار بحرانی توزیع کای اسکور و ۸ درجه آزادی در سطح ۰/۰۱ = α مقایسه شد. نتایج حاصل نشان داد که همه سؤال‌ها به‌جز

چهار سؤال موردمطالعه، به‌طور هم‌زمان DIF و DDF نشان می‌دهند. این تفاوت را می‌توان به این واقعیت نسبت داد که تحلیل بر اساس مدل دو پارامتری لوجیت آشیانی برخلاف رویکرد پاسخ اسمی، آزمونِ گزینه پاسخ صحیح را از آزمونِ گزینه‌های انحرافی جدا می‌کند و لذا DDF کمتری نشان می‌دهد.

در ادامه جهت تفسیر یافته‌ها، تمرکز خود را بر دو سؤالی که تحت هردو آزمون ۱ و ۲ رد شده‌اند و درنتیجه هم DIF و هم DDF نشان دادند (سؤال‌های ۱۳ و ۱۹) قرار می‌دهیم؛ چراکه در این‌گونه سؤال‌ها است که می‌توان منبع DIF را در گزینه‌های انحرافی جستجو کرد. برای بررسی نوع DIF/ DDF در این دو سؤال، یک‌بار با یکسان در نظر گرفتن پارامتر شیب و برآورد آزاد پارامتر دشواری و یک‌بار با یکسان در نظر گرفتن پارامتر دشواری و برآورد آزاد پارامتر شیب، منطبق بر آزمون ۱ و ۲، آماره نسبت درستنمایی بین مدل فشرده (یکسانی همه پارامترها برای دو گروه) و هریک از مدل‌های فوق برآورد شد که نتایج آن در جدول ۸ آمده است. ازآنجاکه استفاده از آماره G2 برای بررسی نوع DIF/ DDF امکان افزایش نرخ خطای نوع اول خانوار گونه از سطح معنی‌داری (۰.۰۱) را برای آزمون هم‌زمان هر دو اثر (دشواری و شیب) به همراه می‌آورد؛ بنابراین با استفاده از تعدیل بنفرونی سطح معنی‌داری ۰.۰۲ برای هر اثر (دشواری و شیب) استفاده شد.

جدول ۸. نوع کنش افتراقی در سؤال‌های ۱۳ و ۱۹ آزمون ریاضی مبتنی بر مدل لوجیت آشیانه‌ای

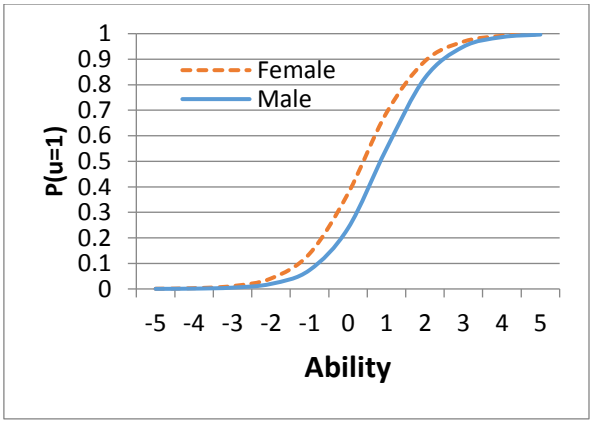
سؤال	آزمون کنش افتراقی گزینه پاسخ				نتیجه	آزمون کنش افتراقی گزینه‌های انحرافی				نتیجه
	دشواری		شیب			عرض از مبدأ		شیب		
	G2	p	*G2*	p		*G2*	p	*G2*	p	
۱۳	۳۳.٦٦	۰.۰۰۰۱>	۱۳.۵۹	۰.۰۰۱>	غیریکنواخت	۸.۱۰	۰.۰٤٤	۷.۳۵	۰.۰٦۱	–
۱۹	۲۱.۹۷	۰.۰۰۰۱>	۰.۰۵	۰.۸۳٤	یکنواخت	۱۰.۱۳	۰.۰۱۷	٦.۷۵	۰.۰۸۰	یکنواخت

همان‌طور که مشاهده می‌شود، مقدار G2 برای آزمون کنش افتراقی سؤال ۱۳ از حیث دشواری به میزان ۳۳.۶۶ و از حیث شیب ۱۳.۵۹ به دست آمد که با توجه به ارزش بحرانی ($\chi^2_{.02}$ =۵/۴۱ ،df = ۱) معنی‌دار است و درنتیجه فرض صفر نبود DIF در پارامتر دشواری و شیب رد می‌شود و نشان می‌دهد که سؤال ۱۳ دارای DIF از نوع غیریکنواخت است. برای سؤال ۱۹ نیز، مقدار G2 برای آزمون کنش افتراقی از حیث دشواری معنی‌دار بود، اما آزمون کنش افتراقی از حیث پارامتر شیب معنی‌دار نشد که نشان می‌دهد سؤال ۱۹ دارای DIF از نوع یکنواخت است. همچنین مطابق جدول ۸، مقدار G2 به‌دست‌آمده برای فرض صفر نبود

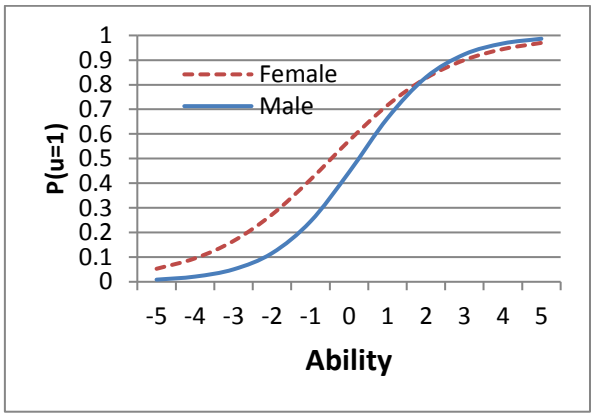
DDF در پارامتر عرض از مبدأ برای هریک از سؤال‌های ۱۳ و ۱۹ به ترتیب ۸.۱۰ و ۱۰.۱۳ به دست آمد که با توجه به ارزش بحرانی (۳ = df، ۹/۸۴ =$\chi^2_{.02}$) نشان می‌دهد که سؤال ۱۳ فاقد DDF درحالی‌که سؤال ۱۹ دارای DDF از نوع یکنواخت است.

منحنی‌های پاسخ صحیح برای این دو سؤال نیز، یک اندازه نسبتاً بزرگ از تفاوت را برای بیشتر سطوح توانایی ارائه می‌دهند به‌طوری‌که احتمال یافتن پاسخ صحیح برای مردان در بیشتر سطوح توانایی سؤال ۱۹ (شکل ۱- ب) و سطوح میانی سؤال ۱۳ (شکل ۱- الف) کمتر از زنان است؛ به‌عبارت‌دیگر سؤال ۱۹ دارای کنش افتراقی (DIF) از نوع یکنواخت به نفع زنان و سؤال ۱۳ دارای کنش افتراقی (DIF) از نوع غیریکنواخت است. بازبینی نمودار احتمال انتخاب گزینه‌های انحرافی برای این سؤال‌ها نیز نشان می‌دهد که بزرگ‌ترین تفاوت‌های مشاهده‌شده بین منحنی‌ها، مربوط به گزینه بدون پاسخ (M4 و F4) این سؤال‌ها است (شکل ۲).

شکل ۱. منحنی‌های ویژه (ICC) پاسخ صحیح سؤال در گروه مردان وزنان در آزمون ریاضی

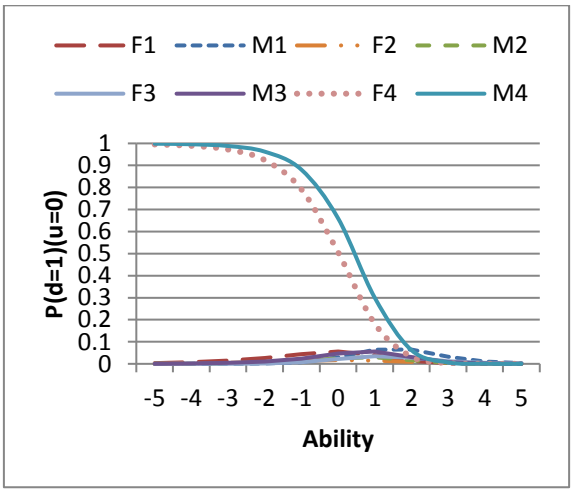


الف- سؤال ۱۳

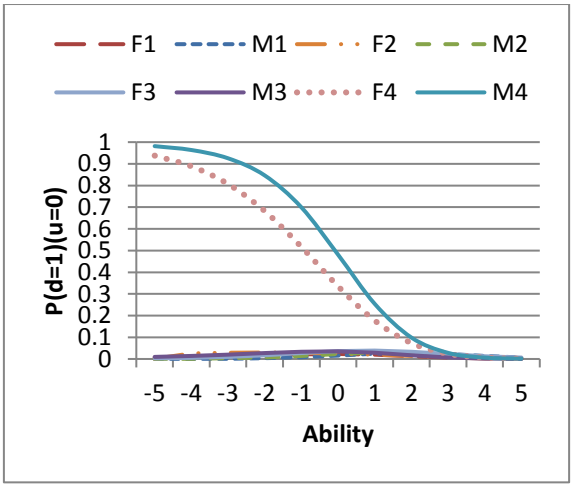


ب- سؤال ۱۹

شکل ۲. منحنی‌های ویژه (ICC) انتخاب گزینه‌های انحرافی برای مردان وزنان در آزمون ریاضی اندازه اثر



الف-سوال ۱۳



ب- سؤال ۱۹

شاخص معیار یابی[1] (STD) دامنه‌ای بین ۱- تا ۱+ اختیار می‌کند، مقادیر بین [۰.۰۵- و ۰.۰۵+] ناچیز شمرده می‌شوند. ارزش‌های بین [۰.۰۵- و ۰.۱۰-] یا [۰.۰۵ و ۰.۱۰] می‌بایست برای اثر ممکن مورد بازبینی قرار گیرد و ارزش‌هایی بزرگ‌تر از قدر مطلق یک (|۱|<) بسیار نامعمول‌اند و باید به‌دقت بررسی شوند (اشمیت و دورانز، ۱۹۹۰). برای کنش افتراقی یکنواخت ارزش‌های منفی این شاخص نشانگر کنش افتراقی علیه گروه کانونی (زنان) و ارزش‌های مثبت نشانگر کنش افتراقی علیه گروه مرجع (مردان) است. برای کنش افتراقی غیریکنواخت از اندازه‌های بدون علامت بر اساس قدر مطلق تفاوت استفاده می‌شود (Suh & Talley, 2015). برای نمونه اندازه اثر دو سؤال که به‌عنوان DDF/DIF دار شناسایی شد، در جدول ۹ آمده است. علت تمرکز بر این سؤال‌ها آن بود که این سؤال‌ها نشانگر پتانسیل بالقوه گزینه‌های انحرافی در ایجاد DIF هستند.

جدول ۹. اندازه‌های حجم اثر برای DDF/DIF سؤال‌های ۱۳ و ۱۹ آزمون ریاضی

شاخص	کنش افتراقی		اندازه اثر		بزرگی اثر	
			سؤال ۱۳	سؤال ۱۹	سؤال ۱۳	سؤال ۱۹
STD	DIF	علامت‌دار	۰/۰۸۹	۰/۰۹٤	متوسط	بزرگ
		بدون علامت	۰/۰۸۹	۰/۱۰۷		
	DDF	علامت‌دار	۰/۰٤۹ (گزینه ۵)	۰/۰٤۷- (گزینه ۵)	متوسط	تقریباً متوسط
		بدون علامت	۰/۰۹۰(گزینه ۵)	۰/۰٤۷(گزینه ۵)		

1. Standardization Index

شاخص	کنش افتراقی		اندازه اثر		بزرگی اثر	
			سؤال ۱۳	سؤال ۱۹	سؤال ۱۳	سؤال ۱۹
		علامت‌دار	۰/۰۰۶(گزینه ۳)	۰/۰۲۰(گزینه ۳)	جزئی	جزئی
		بدون علامت	۰/۰۱٤(گزینه ۳)	۰/۰۳۰(گزینه ۳)		
		علامت‌دار	۰/۰۰۵(گزینه ۲)	۰/۰۳۱(گزینه ۲)	جزئی	جزئی
		بدون علامت	۰/۰۱٦(گزینه ۲)	۰/۰۳۱(گزینه ۲)		
		علامت‌دار	۰/۰۲۸(گزینه ۱)	–	جزئی	–
		بدون علامت	۰/۰۳٤(گزینه ۱)	–		

* گزینه ۵ مربوط به افرادی است که به سؤال پاسخ نداده‌اند

مطابق جدول ۹ برای DIF در سؤال ۱۳ اندازه‌های اثر علامت‌دار و بدون علامت، ارزش‌های یکسانی به میزان ۰/۰۸۹ ارائه می‌دهد که نشانگر یک DIF متوسط و یکنواخت علیه گروه مرجع (مردان) است. برای DIF سؤال ۱۹، مقادیر علامت‌دار و بدون علامت، یکسان نیستند و مقدار بدون علامت به میزان جزئی بیشتر از مقدار علامت‌دار است که نشان می‌دهد یک شانس جزئی برای وجود DIF غیریکنواخت قابل‌تصور است. به‌هرحال، سؤال ۱۹ یک DIF در حد بزرگ (۰/۱۰۷) را نشان می‌دهد. برای DDF در سؤال ۱۳ اندازه‌های اثر علامت‌دار و بدون علامت بیانگر DDF غیریکنواخت است. در این میان بزرگ‌ترین اثر بدون علامت متعلق به گزینه ۵ (گزینه بدون پاسخ) و به میزان ۰/۰۹ به دست آمد که بیانگر کنش افتراقی متوسط است. همچنین سایر گزینه‌های انحرافی این سؤال، اثری جزئی را به نمایش گذاشت. DDF در سؤال ۱۹ اثراتی جزئی برای دو گزینه ۲ و ۳ به نفع گروه کانونی (زنان) نشان داد. در این میان بزرگ‌ترین اثر متعلق به گزینه ۵ به میزان ۰/۰۴۷- بود که با توجه به علامت و اندازه آن بیانگر یک DDF تقریباً متوسط به نفع گروه مرجع (مردان) است.

بحث و نتیجه‌گیری

هدف از این مطالعه، ارزیابی رویکرد لوجیت آشیانه‌ای دو پارامتری در شناسایی سؤال‌های DIF / DDF دار با استفاده از داده‌های شبیه‌سازی و داده‌های واقعی بود. بر اساس تحلیل داده‌های شبیه‌سازی و به‌منظور پاسخگویی به سؤال اول پژوهش مبنی بر: «آیا شناسایی کارکرد افتراقی گزینه‌های انحرافی با استفاده از آزمون LR مدل لوجیت دو پارامتری آشیانه‌ای تحت تأثیر توزیع توانایی (یکسان و متفاوت)، ویژگی‌های سؤال (دشواری و

تشخیص)، اندازه‌ی DDF (متوسط، بزرگ) و اندازه‌ی DIF (کم، متوسط، زیاد) قرار می‌گیرد؟»؛ نتایج زیر حاصل شد:

– متوسط تعداد سؤال‌هایی که تحت شرایط مختلفِ شبیه‌سازی به‌درستی به‌عنوان DIF دار، یا به‌عنوان DDF دار شناسایی شد به ترتیب ۸۷/۷٪ و ۹۷/۵٪، بود که نشان می‌دهد رویکرد لوجیت آشیانه‌ای از توان خوبی بخصوص در ارتباط با DDF برخوردار است.

– در نبود هرگونه کنش افتراقی و تحت شرایط توزیع یکسان توانایی (نبود اثر)، متوسط نرخ خطای نوع اول برای آزمون DIF و DDF به ترتیب ۰/۰۴۳ و ۰/۰۴۷ به دست آمد که در محدوده آلفای اسمی قرارداد، هرچند تحت توزیع نابرابر توانایی، مقداری تورم نرخ خطا (به ترتیب ۰/۰۵۵ و ۰/۰۵۴) یافت شد. چنین نتایجی را می‌توان در سایر مطالعات مشابه نیز یافت (برای نمونه؛ Suh & Bolt, 2011؛ Penfield, 2008). چنین نتایجی قابل‌انتظار بود چراکه معلوم شده بسیاری از آماره‌های DIF هنگامی‌که توزیع‌های توانایی متفاوت باشد، دچار تورم نرخ خطای نوع اول می‌شوند. (Judin & Girel, 2001؛ Narayanan & Swaminathan, 1996؛ Penfield, 2010، Guler, & Penfield, 2009).

– نرخ شناسایی کنش افتراقی سؤال متأثر از پارامترهای سؤال (دشواری و شیب) بود؛ به شکلی که با افزایش دشواری سؤال‌ها و یا کاهش شیب، قدرت شناسایی کنش افتراقیDIF افزایش می‌یافت. چنین روابطی در دیگر مطالعات نیز مشاهده‌شده است (Suh & Bolt, 2011؛ Hidalgo & López-Pina, 2004). هرچند به نظر می‌رسد این نتایج تا حدودی بستگی به‌دشواری سؤال دارد، به‌طوری‌که با دشوارتر شدن سؤال، اختلاف بین گروه مرجع و گروه اقلیت بیشتر نمود می‌یابد، ممکن است ناشی از این واقعیت هم باشد که با افزایش پارامتر شیب، مساحت بین ICC های مرتبط با تفاوت گروهی در پارامتر a کاهش می‌یابد (Hidalgo & López-Pina, 2004). همچنین هماهنگ با نتایج دیگر محققان (مثلاً Suh & Bolt, 2011) یافته‌ها نشان داد نرخ شناسایی کنش افتراقی گزینه‌های انحرافی نیز متأثر از پارامترهای سؤال (دشواری و قوه تشخیص) است. به‌طوری‌که با کاهش سطح دشواری و یا با افزایش قوه تمیز سؤال، توان تشخیص صحیح DDF افزایش می‌یابد. چنین نتایجی قابل‌انتظار است، چراکه با افزایش قدرت تمیز سؤال، پاسخ‌های غلط برای افرادِ با توانایی اندک، جذابیت بیشتری پیدا می‌کند.

– صرف‌نظر از توزیع توانایی گروه‌ها، به‌طور متوسط با افزایش سطح DIF و یا افزایش سطح DDF –به‌استثنای چند مورد که ممکن است ناشی از خطای نمونه‌گیری باشد– نرخ رد فرض صفر افزایش می‌یابد. این نتایج با مطالعات Suh and Bolt (2011) کاملاً سازگار است و به شکلی (در رابطه با DIF) در تحقیقات دیگران نیز مشاهده‌شده است (Hidalgo & López-Pina, 2004).

بر اساس تحلیل داده‌های واقعی و به‌منظور پاسخگویی به سؤال دوم پژوهش مبنی بر: «چه رابطه‌ای بین کارکرد افتراقی سؤال و کارکرد افتراقی گزینه‌های انحرافی در آزمون‌های واقعی وجود دارد؟» یافته‌ها نشان داد که زمانی که در رابطه با پاسخ صحیح،DIF وجود دارد،DDF می‌تواند رخ دهد یا رخ ندهد. چنین نتیجه‌ای قابل‌تصور بود چراکه اساساً روش لوجیت آشیانه‌ای یک رویکرد مبتنی بر چارچوب «تقسیم‌بر گزینه‌های انحرافی» است که تفکیک دو علت DIF را با بکار بردن یک استراتژی دومرحله‌ای امکان‌پذیر می‌سازد (Suh & Bolt, 2011).

بر اساس نتایج تحلیل تجربی با داده‌های واقعی، ۹ سؤال از ۱۳ سؤال موردمطالعه DIF نشان داد، ۴ سؤال به‌عنوان DDF دار و ۲ سؤال به‌طور هم‌زمان هردوی DIF و DDF را به نمایش گذاشت، درحالی‌که در رویکرد رقیبِ پاسخ اسمی، ۱۱ سؤال به‌عنوان سؤال با کنش افتراقی شناسایی شد؛ بنابراین همان‌طور که انتظار می‌رفت رویکرد لوجیت آشیانه‌ای مبتنی بر استراتژی «تقسیم‌بر گزینه‌های انحرافی» نسبت به رویکرد مبتنی بر استراتژی «تقسیم‌بر کل»، تعداد سؤال‌های کمتری را به‌عنوان DDF دار شناسایی نمود که اشاره بر محافظه‌کاری بیشتر این رویکرد در تشخیص DDF دارد (Suh & Talley, 2015).

بازبینی نمودار احتمال انتخاب گزینه‌های انحرافی برای دو سؤال DIF دارِ ۱۳ و ۱۹ نشان داد که بزرگ‌ترین تفاوت‌های مشاهده‌شده بین منحنی‌ها در گزینه «بدون پاسخ» رخ‌داده است. ممکن است این مسئله یک علت بروز DIF بخصوص در سؤال ۱۹ باشد؛ به‌عبارت‌دیگر ممکن است ریسک‌پذیری بیشتر پسران نسبت به دختران دردادن پاسخ‌های حدسی سبب کنش افتراقی سؤال به نفع دختران گردیده باشد. به‌طورکلی بررسی الگوی پاسخ‌های انحرافی نشان داد که علت اثر DIF در برخی سؤال‌ها ممکن است ناشی از وجود عامل سوگیری در گزینه پاسخ و یا تنه سؤال (سؤال‌های واجدDIF و فاقد DDF) و در برخی دیگر ناشی از کارکرد گزینه‌های انحرافی (مثلاً سؤال ۱۹) باشد.

ازآنجاکه توان آماری یک آزمون به حجم نمونه وابسته است، بررسی اندازه اثر نقش مهمی در تعیین کنش افتراقی سؤال به‌ویژه هنگامی‌که حجم نمونه بزرگ است، دارد. در این مطالعه کاربرد اندازه‌های اثر، اطلاعات توصیفی بیشتری چون: نوع، جهت و بزرگی کنش افتراقی را فراهم نمود. علاوه بر آن اندازه اثر برای هر طبقه پاسخ نشان داد کدام طبقه پاسخ به‌طور بالقوه مسبب DIF است.

در سؤال‌های چندگزینه‌ای، کنش افتراقی سؤال (DIF) از حیث طبقه پاسخ صحیح، ممکن است ناشی از کارکرد افتراقی گزینه‌های انحرافی باشد یا نباشد (Suh & Talley, 2015). تعیین گزینه‌های انحرافی به‌عنوان علل DIF، می‌تواند اطلاعات باارزشی برای تجدیدنظر احتمالی در سؤال یا طراحی سؤال‌های جدید فراهم نماید. رویکرد دومرحله‌ای مبتنی بر کاربرد مدل لوجیت آشیانه‌ای دو پارامتری، ضمن تفکیک آزمون کنش افتراقی گزینه‌های انحرافی (DDF) از کنش افتراقی سؤال (DIF)، امکان ارزیابی روشن از اینکه آیا گزینه‌های انحرافی مسئول احتمالی DIF هستند را میسر می‌سازد. این ابزار به فهم علت DIF و تعیین موقعیت عامل سوگیری کمک می‌کند، ضمن آنکه شواهدی را برای کارشناسان محتوا در تشخیص گزینه‌های مشکل‌دار فراهم می‌آورد. ازآنجاکه آزمون‌های سرنوشت‌ساز نقش مهمی در گزینش افراد دارد و تحلیل‌های DIF و DDF جایگاه ویژه‌ای در تعیین اعتبار و نامتغیر بودن اندازه‌گیری سؤال‌های این آزمون‌ها دارند، پیشنهاد می‌شود جهت سرند کردن سؤال‌های سودار در آزمون‌های چندگزینه‌ای تحلیل‌های جامع DIF / DDF موردتوجه قرار گیرد.

همانند دیگر مطالعات مونت‌کارلو، تعداد شرایط قابل دست‌کاری در این مطالعه محدود بود تا مدیریت مطالعه میسر شود بنابراین پیشنهاد می‌شود در مطالعات شبیه‌سازی‌شده آتی متغیرهایی چون طول آزمون، تعداد گزینه‌ها و حجم نمونه متفاوت برای گروه‌های کانونی و مرجع نیز موردبررسی قرار گیرد. هرچند در پژوهش حاضر کنش افتراقی از نوع غیریکنواخت موردمطالعه قرار گرفت، اما پارامتر شیب در همه شرایط یکسان (۰.۳) بود. ازآنجاکه تغییر پارامتر شیب در سؤال‌های دارای کنش افتراقی می‌تواند بر توان ردگیری DIF یا احتمالاً DDF تأثیر بگذارد، نتایج حاصل از این مطالعه در رابطه با کنش غیریکنواخت را نمی‌توان به دیگر سؤال‌های کنش دار باارزش‌های مختلف پارامتر شیب تعمیم داد. لذا مطالعات شبیه‌سازی‌شده بیشتری باید صورت گیرد که در آن نه‌تنها اثر تغییر

پارامتر شیب، بر کارکرد افتراقی غیریکنواخت، بلکه کنش افتراقی یکنواخت نیز موردبررسی قرار گیرد. در این مطالعه برای بررسی کنش افتراقی، زمانی که توزیع توانایی گروه‌ها نابرابر است (وجود اثر) از اختلافی به میزان ۰/۵ واحد استاندارد استفاده شد، پیشنهاد می‌شود در مطالعات آتی سطوح دیگر اختلاف توزیع توانایی نیز بررسی شود تا تصویر بهتری از عملکرد روش‌های ردیابی کنش افتراقی به دست آید.

منابع

امبرتسون، سوزان ای. رایس، استیونس پی. (۲۰۰۰). نظریه‌های جدید روان‌سنجی برای روانشناسان./ ترجمه حسن پاشا شریفی و همکاران، ۱۳۸۸. تهران: رشد.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Banks, Kathleen. (2009). Using DDF in a Post Hoc Analysis to Understand Sources of DIF. *Educational Assessment*, 14:103–118

Bock, R. D. (1972) Estimating item parameters and latent proficiency when the responses are scored in two or more nominal categories. *Psychometrika, 37*., 29-51.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Sage Publications: Thousand Oaks, California.

DeMars, C. E. (2010). Type I error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement*, *70*, 961–972

Dorans, N. J and. Cook, Linda L (2016), *Fairness in Educational Assessment and Measurement.* NY: Taylor & Francis.

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement, 29*(4), 309-319.

Douglas, J., Roussos, L., & Stout, W. (1996). tem bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF. *Journal of Educational Measurement, 33*, 465–484.

Feinberg, R. A. and Rubright, J. D. (2016). Conducting Simulation Studies in Psychometrics. *Educational Measurement: Issues and Practice.* Summer, Vol. 35, No. 2, pp. 36–49

Green, B. F., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distractor functioning. *Journal of Educational Measurement, 26*(2), 147-160.

Guler, N. & Penfield, R. D. (2009). A Comparison of the Logistic Regression and Contingency Table Methods for Simultaneous Detection of Uniform and Nonuniform DIF. *Journal of Educational Measurement.* Fall, Vol. 46, No. 3, pp. 314–329

Harwell, M., Stone, C.,A., Hsu, Tse-Chi and Kirisci, L. (1996). Monte Carlo Studies in Item Response Theory. *Applied Psychological Measurement*. Vol. 20, No. 2, June, 101-125

Holland, P. W., & Thayer, D. T. (Eds.). (1988). *Differential item performance and theMantel-Haenszel procedure.* Hillsdale, NJ: Erlbaum.

Hidalgo, M.D., López-Pina, J.P. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel–Haenszel procedures. *Educational and Psychological Measurement*, 64, 903–915.

Hutchinson, T.P. (1991). Ability, partial information, and guessing: *Statistical modeling applied to multiple-choice tests.*Rundle Mall: Rumsby Scientific.

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, 329-349.

Kato, K., Moen, R. E., & Thurlow, M. L. (2009). Differentials of a State Reading Assessment: Item Functioning, Distractor Functioning, and Omission Frequency for Disability Categories. *Educational Measurement: Issues and Practice*.Volume 28, Issue 2, Summer, Pages 28–40.

Kim, Seock-Ho & Cohen, Allan S. (1995). A Comparison of Lord's Chi-Square, Raju's Area Measures, and the Likelihood Ratio Test. on Detection of Differential Item Functioning, *Applied Measurement in Education*, 8:4, 291-312.

Koon, S. Kamata A. (2013). An applied examination of methods for detecting differential distractor functioning. *International Journal of Quantitative Research in Education*. Vol.1 No.4

Li,Z. (2014). Power and Sample Size Calculations for LogisticRegression Tests for Differential Item Functioning. *Journal of Educational Measurement.Winter 2014, Vol. 51, No. 4, pp. 441–462*

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Middleton K. and Laitusis C. Cahalan (2007). Examining Test Items for Differential Distractor Functioning Among Students With Learning Disabilities. (Research Report) ETS, Princeton, NJ

Mapuranga, R. Dorans, Neil J. and Middleton K. (2008). A Review of Recent Developments in Differential Item Functioning. (Research Report) ETS, Princeton, NJ

Narayanan, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF.*Applied Psychological Measurement, 20*, 257–274.

Penfield Randall D. and Camilli Gregory (2007). Differential Item Functioning and Item Bias. In C.R. Rao & S. Sinharay (Eds.), *Handbook of Statistics on Psychometrics, Vol. 26.* (pp.125-167). Elsevier B.V.

Penfield, R. D. (2008). An odds ratio approach for assessing differential distractor functioning effects under the nominal response model. *Journal of Educational Measurement*, 45, 247-269.

Penfield, R. D. (2010). Modeling DIF Effects Using Distractor-Level Invariance Effects: Implications for Understanding the Causes of DIF. *Applied Psychological Measurement*, 34(3) 151–165

Penfield, R. D. (2016). Fairness in Test Scoring. In Neil J. Dorans and Linda L. Cook (Eds.), *Fairness in Educational Assessment and Measurement.* (pp.125-167). NY: Taylor & Francis.

Reif, M. (2015). *mcIRT: Software Package for multiple choice items IRT models*. Available on https://github.com/manuelreif/mcIRT

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel Type I error performance. *Journal of Educational Measurement, 33*, 215–230.

Schmitt, A. P., & Bleistein, C. A. (1987). Factors affecting differential item functioningfor black examinees on Scholastic Aptitude Test analogy items. *Research report no. 87-23.* Educational Testing Service, Princeton, NJ.

Schmitt, A. P., & Dorans, N. J. (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement, 27*(1), 67-81.

Schmitt, A. P., Holland, P.W., & Dorans, N. J. (1993). Evaluating hypotheses about differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281–315). Hillsdale, NJ: Lawrence Erlbaum.

Suh, Y. & Talley, A. E. (2015). An Empirical Comparison of DDF Detection Methods for Understanding the Causes of DIF in Multiple-Choice Items. *Applied Measurement in Education*, 28: 48–67, 2015

Suh, Y., & Bolt, D. M. (2010). Nested logit models for multiple-choice item response data. *Psychometrika*, *75*, 454–473.

Suh, Y., & Bolt, D. M. (2011). A nested logit approach for investigating distractors as causes of differential item functioning. *Journal of Educational Measurement*, *48*, 188–205

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*, 501-519.

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-choice items: The distractors are also part of the item. *Journal of Educational Measurement, 26*, 161-176.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Wang, W.-C. (2000). Factorial modeling of differential distractor functioning in multiple-choice items. *Journal of Applied Measurement*, *I*, 238-256.

Zieky, M. (2006). Fairness reviews in assessment. In S. M. Downing & T. M. Haladyna (Eds.) *Handbook of test development* (pp. 359–376). Mahwah, NJ: Lawrence Erlbaum Associates