Iranian Journal of Psychoeducational Assessment
University of Zanjan

January 2026. Vol 1. Issue 2

Research Paper

The Effect of Discrimination and Difficulty Parameters on the Efficiency of Differential Item Functioning Approaches in the Presence of a Guessing Factor

Ahmad Rezaei[1*], **Mohammadreza Falsafinejad** [2], **Masoud Geramipour**, [3]

1. Assistant Professor, Special Operations Department, Faculty of Police Sciences and Technologies, Police University, Tehran, Iran
2. Associate Professor, Assessment and Measurement Department, Faculty of Psychology and Educational Sciences, Allameh Tabataba'i University, Tehran, Iran
3. Associate Professor, Curriculum Planning Studies Department, Faculty of Psychology and Educational Sciences, Kharazmi University, Tehran, Iran

Article info:
Received: 05.06.2025
Revised: 18.09.2025
Accepted: 09.10.2025

Keywords:
psychometrics, differential item functioning, guess effect, simulation, item parameters

Publisher: University of Zanjan

Abstract

Measurement bias poses a serious threat to test validity, and research has consistently shown that item parameters play a crucial role in generating such bias. The present study aimed to identify the optimal three-parameter method for detecting differential item functioning (DIF) while considering the effects of item parameters. Designed as an applied psychometric investigation, the study simulated data based on independent variables including test type (nonlinear regression, Lord's three-parameter, and Raju's three-parameter methods), ranges of parameters (low, medium, and high), and item characteristics. Simulation was conducted by manipulating 20% of items to exhibit DIF with an effect size of 0.5, producing 10 sets of data and a total of 90 datasets. For each dataset, three DIF detection tests were performed, resulting in 270 analyses overall. Findings revealed that the main effect of test type was statistically significant, with nonlinear regression outperforming Raju's three-parameter method in terms of accuracy. Moreover, DIF detection was more effective when the difficulty parameter was at a medium level compared to low or high levels, while increases in the discrimination parameter consistently enhanced correct detection rates. In contrast, variations in the guessing parameter did not yield significant differences, as test performance remained stable across all levels of guessing. These results underscore the importance of focusing on difficulty and discrimination parameters in DIF detection and suggest that nonlinear regression can serve as a practical and efficient alternative to item response theory (IRT)-based methods, particularly in contexts where guessing effects are present. Overall, the study contributes to psychometric methodology by highlighting parameter-specific influences and offering evidence for the utility of nonlinear regression in improving DIF detection.

Use your device to scan and read the article online

Citation: Rezaee, A., Falsafinejad, M. & Gharamipour, M. (2026). The Effect of Discrimination Parameters on the Efficiency of Differential Functioning Approaches in the Presence of a Guessing Factor. *Iranian Journal of Psychoeducational Assessment*, 1 (2), 223-236. https://doi.org/10.30470/ijpa.2025.733215

***Corresponding Author:** Ahmad Rezaei
Address: Special Operations Department, Faculty of Police Sciences and Technologies, Police University, Tehran, Iran
Email: amdrezaee@yahoo.com

Extended Abstract

Introduction

Accurate and valid measurement is fundamental to scientific research, particularly in psychology and educational sciences, where tests and questionnaires inform important educational and professional decisions. A key condition for test validity is fairness in measurement, meaning that individuals with the same level of ability should have equal chances of responding correctly to test items regardless of group membership such as gender, ethnicity, or social background. Differential Item Functioning (DIF) threatens this fairness by arising when an item favors one group over another despite equivalent ability levels. Identifying and eliminating items with DIF is therefore essential, especially in high-stakes assessments, to ensure that decisions based on test scores are equitable and unbiased. Detecting DIF, however, is a complex process influenced by several factors. Item characteristics and psychometric parameters particularly item difficulty, discrimination, and guessing can substantially affect the accuracy of DIF detection methods. For example, extremely easy or difficult items may reduce the sensitivity of statistical procedures. Traditional and widely used methods, such as the Mantel–Haenszel procedure and logistic regression, have the advantage of simplicity but generally ignore the guessing parameter, which can lead to biased results. In contrast, more advanced Item Response Theory (IRT)-based approaches, such as Lord's and Raju's methods, explicitly model guessing and provide more precise detection but often require large sample sizes and involve complex computations, limiting their practical application. As an alternative, nonlinear regression has been proposed as a score-based DIF detection method inspired by the three-parameter IRT model. This approach accounts for guessing by modeling a non-zero lower asymptote in the item characteristic curve and may offer a more practical balance between statistical rigor and feasibility. Despite its theoretical promise, there is limited empirical evidence comparing its performance with established IRT-based methods under varying item conditions. Therefore, the present study aims to compare the effectiveness of three three-parameter DIF detection methods (Lord's method, Raju's method, and nonlinear regression) using simulation data, and examine.

Methods

This study is a simulation study in the field of psychometrics. In this study, the independent variables include the DIF detection method (Lord's, Raju's, and Nonlinear Regression), the type of item parameter (difficulty, discrimination, and guessing), and the level of the parameter (low, medium, and high). The dependent variable is the correct DIF detection rate, defined as the proportion of items with DIF that are correctly identified by each method. Test response data were simulated using Wingen3 software under the three-parameter logistic model with a standard normal ability distribution. In each dataset, 20% of the items were designed to contain DIF. By combining different levels of item parameters, 90 datasets were generated, each replicated ten times. DIF analyses were then conducted in R using the difNLR function for nonlinear regression and the difLord and difRaju functions for Lord's and Raju's methods, respectively. The correct DIF detection rate was calculated for each method, and the results were compared using a three-way ANOVA followed by Tukey's post hoc test in SPSS.

Results

The findings showed that all three independent variables (method type, parameter type, and parameter level) had significant effects on DIF detection accuracy, and the interaction between parameter type and parameter level was also significant. The three methods demonstrated clearly different performances: Nonlinear Regression achieved the highest accuracy and

significantly outperformed Raju's method, while Lord's method also surpassed Raju's, though its advantage over Nonlinear Regression was not statistically significant, indicating that both Nonlinear Regression and Lord's are superior choices. Item parameters influenced performance as well, with all methods working most effectively when difficulty and discrimination were at medium levels; extremely easy or hard items, and items with very low or very high discrimination, reduced accuracy, consistent with earlier research. The guessing parameter did not show a consistent overall effect, though Lord's and Raju's methods improved slightly as guessing increased, suggesting that the relationship between guessing and DIF is more complex than previously assumed.

Discussion

The findings of this simulation study suggest that the Nonlinear Regression method can be an efficient and accurate option for DIF detection, particularly when there is concern about the "guessing" phenomenon. By combining the relative simplicity of score-based methods with the capability to model guessing, this method demonstrated competitive or superior performance compared to traditional IRT methods and can therefore be considered a practical and powerful alternative in the toolkit of test analysts. The results clearly confirm that designing items with medium difficulty and discrimination is not only desirable from a psychometric standpoint but also maximizes the accuracy of statistical methods in identifying bias. This finding carries an important practical message for test developers and analysts, emphasizing the necessity of attending to the psychometric quality of items during the design phase, prior to engaging in complex DIF analysis.

As a limitation, this study was conducted within a simulated data environment. Future research is recommended to investigate the performance of these methods using real data from high-stakes tests (e.g., university entrance exams) and under more complex conditions (e.g., non-uniform DIF). In summary, this research is a step towards clarifying the roadmap for selecting an optimal DIF detection method. Awareness of the influence of item parameters assists measurement specialists in choosing appropriate analytical methods and ultimately contributes to the design and use of fairer and more valid assessments.

Ethical considerations

Since the present study utilized simulated data, no human or animal participants were involved, and no ethical concerns were raised.

Funding

This study was conducted without any external funding, and all related costs were personally borne by the authors.

Authors' contributions

MG contributed to the topic selection and theoretical framework, MF to the research methodology, and AR to the implementation and initial drafting of the manuscript. The final version of the article was reviewed and approved by all three authors.

Conflict of interest

The authors declare no conflict of interest.

مقاله پژوهشی

اثر پارامترهای تمیز و دشواری در کارایی رویکردهای تشخیص کنش افتراقی سؤال با حضور عامل حدس

احمد رضائی[1*]، محمدرضا فلسفی‌نژاد[2]، مسعود گرامی‌پور[3]

۱. استادیار گروه عملیات ویژه، دانشکده علوم و فنون انتظامی، دانشگاه پلیس، تهران، ایران
۲. دانشیار گروه سنجش و اندازه‌گیری، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه علامه طباطبائی، تهران، ایران
۳. دانشیار گروه آموزشی مطالعات برنامه‌ریزی درسی، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه خوارزمی، تهران، ایران

اطلاعات مقاله:
تاریخ دریافت: ۱۴۰۴/۰۳/۱۵
تاریخ داوری: ۱۴۰۴/۰۶/۲۷
تاریخ پذیرش: ۱۴۰۴/۰۷/۱۷

واژه‌های کلیدی:
روانسنجی، کنش افتراقی سؤال، اثر حدس، پارامترهای سؤال، رگرسیون غیرخطی

ناشر: دانشگاه زنجان

چکیده

سوگیری اندازه‌گیری می‌تواند روایی آزمون را به خطر بیندازد و یافته‌ها نشان می‌دهد پارامترهای سؤال در ایجاد سوگیری آزمون نقش مهمی دارند. هدف پژوهش حاضر شناسایی روش بهینه سه‌پارامتری برای تشخیص کنش افتراقی سؤال با در نظر گرفتن اثر پارامترهای سؤال بود. این تحقیق از نوع کاربردی و در زمره مطالعات روانسنجی قرار می‌گیرد. داده‌های مورد نیاز بر اساس متغیرهای مستقل شامل نوع آزمون (رگرسیون غیرخطی، لرد و راجو سه‌پارامتری)، طیف پارامترها (کم، متوسط و زیاد) و پارامترهای سؤال شبیه‌سازی شد. شبیه‌سازی با دستکاری ۲۰ درصد سؤالات دارای کنش افتراقی با شدت اثر ۰/۵ انجام گرفت. داده‌ها در ۱۰ مجموعه و در مجموع ۹۰ مجموعه تولید شد و برای هر مجموعه سه آزمون تشخیص کنش افتراقی اجرا شد که در مجموع ۲۷۰ آزمون را شامل می‌شد. نتایج نشان داد اثر اصلی متغیر آزمون معنادار است و آزمون رگرسیون غیرخطی عملکرد بهتر و معناداری نسبت به آزمون راجو سه‌پارامتری دارد. همچنین، در شرایطی که پارامتر دشواری در سطح متوسط قرار دارد، میزان تشخیص صحیح کنش افتراقی بیشتر از شرایطی است که این پارامتر در سطح کم یا زیاد باشد. علاوه بر این، افزایش پارامتر تمیز باعث افزایش میزان تشخیص صحیح کنش افتراقی توسط آزمون‌ها شد. در مقابل، طیف‌های مختلف پارامتر حدس، تفاوت معناداری در کارایی آزمون‌ها ایجاد نکردند و عملکرد آزمون‌ها در تمام سطوح این پارامتر یکسان بود. این یافته‌ها نشان‌دهنده اهمیت توجه به پارامترهای دشواری و تمیز در تشخیص کنش افتراقی سؤال است. بر اساس نتایج، می‌توان نتیجه گرفت که روش رگرسیون غیرخطی در آزمون‌هایی که اثر حدس وجود دارد، می‌تواند به‌عنوان جایگزینی مفید و کارآمد برای روش‌های مبتنی بر نظریه سؤال–پاسخ (IRT) در تشخیص کنش افتراقی سؤال مورد استفاده قرار گیرد.

استناد: رضائی، ا.، فلسفی‌نژاد، م. و گرامی‌پور، م. (۱۴۰۴). اثر پارامترهای تمیز و دشواری در کارایی رویکردهای تشخیص کنش افتراقی سؤال با حضور عامل حدس. *سنجش روانی تربیتی*، ۱(۲)، ۲۲۳–۲۳۶.
https://doi.org/10.30470/ijpa.2025.733215

* **نویسنده مسئول:** احمد رضائی
نشانی: دانشکده علوم و فنون انتظامی، دانشگاه پلیس، تهران، ایران
پست الکترونیکی: amdrezaee@yahoo.com

مقدمه

اساس علم، اندازه‌گیری است و توسعه ابزارهای قابل اعتماد و معتبر از دیرباز برای پژوهشگران اهمیت ویژه‌ای داشته است (بائر[1] و همکاران، ۲۰۲۰). پیشرفت هر علمی تا حد زیادی به در دسترس بودن ابزارهای دقیق و معتبر برای اندازه‌گیری بستگی دارد (بائر، ۲۰۱۷). اندازه‌گیری در علوم اجتماعی عمدتاً به‌وسیله آزمون‌ها انجام می‌گیرد، اما آزمون‌ها همیشه واجد ویژگی‌های مطلوب نیستند. همزمانی حضور تمامی مؤلفه‌های لازم برای ایجاد شرایط مطلوب در یک آزمون، وضعیتی ایده‌آل است و انتظار وجود تمامی عوامل ضروری در مسیر تصمیم‌سازی و تصمیم‌گیری برای دستیابی به نتیجه‌ای حقیقی دشوار به نظر می‌رسد. بدیهی است که توجه به برخی از این عوامل اهمیت فراوانی دارد، زیرا چشم‌پوشی از آنها منجر به پایمال شدن حقوق برخی از ذی‌نفعان آزمون می‌شود. در همین راستا، علم روان‌سنجی به دنبال به حداکثر رساندن کارایی آزمون‌ها برای دستیابی به اهداف مورد نظر است (باقری‌خواه و همکاران، ۱۳۹۰).

روایی از جمله ویژگی‌های مهم و ضروری برای یک آزمون است. یکی از جنبه‌های کلیدی روایی این است که آیا نمرات تولید شده توسط یک مقیاس به‌طور مستقیم در افراد مختلف قابل مقایسه است یا خیر. ارزیابی اینکه آیا نمرات از نظر معنا و اندازه‌گیری در افراد مختلف معادل هستند، وضعیتی است که به آن تغییرناپذیری اندازه‌گیری گفته می‌شود (بائر و همکاران، ۲۰۲۰). در صورت وجود تغییرناپذیری اندازه‌گیری، می‌توان نمرات و سایر نتایج را به‌طور معتبر بین افراد مختلف مقایسه کرد. از این رو، روان‌شناسان بر اهمیت همگنی اندازه‌گیری به‌عنوان پیش‌نیازی برای مقایسه‌های گروهی تأکید کرده‌اند (راجو[2] و همکاران، ۲۰۰۲؛ ریس[3] و همکاران، ۱۹۹۳؛ وندربرگ[4] ، ۲۰۰۲). فقدان همگنی اندازه‌گیری و کنش افتراقی سؤال به‌عنوان تهدیدی جدی برای روایی آزمون تلقی می‌شود، به‌طوری که استانداردهای انجمن روان‌شناسی آمریکا[5] (APA) بر بررسی کنش افتراقی سؤال‌های آزمون برای ارزیابی منصفانه بودن آن تأکید دارد (برن و استوارت[6] ، ۲۰۰۶). میلسپ[7] (۲۰۱۱) خاطرنشان می‌کند که سوداری سؤال، که با روایی آزمون ارتباط مستقیم دارد، می‌تواند با سنجش سازه‌ای که برای آن طراحی نشده، معنای آزمون را تغییر دهد. در این شرایط، تصمیمات منطقی نبوده و به‌کارگیری چنین آزمون‌هایی به نتایج نامعقول و غیردقیق می‌انجامد (زومبو[8] ، ۱۹۹۹). بنابراین، مدعیان بسیاری از جمله حقوق شهروندی، قانون آموزش معلولین و قانون حقوق مدنی که مفهوم عدالت را بر اساس نژاد، رنگ، مذهب، جنس و منشأ ملیت پوشش می‌دهند، در برابر آن موضع‌گیری کرده‌اند (کامیلی[9] ، ۲۰۰۶). در پاسخ به ادعای منتقدینی که معتقدند سؤال سودار با داشتن خطای نظام‌مند باعث نامعتبر شدن نتایج سنجش می‌شود، مجریان ملزم هستند نشان دهند که آزمون‌هایشان در برابر اقلیت‌ها عاری از سوگیری است (همبلتون[10] و همکاران، ۱۳۸۹؛ آنگوف[11]، ۱۹۹۳). از این رو، روان‌سنج‌هایی مانند کریستیانسن، ورس، مکدوال و زومبو (۲۰۰۵) و کامیلی و شپارد (۱۹۹۴) بر این موضوع هم‌عقیده‌اند که شناسایی سؤال سودار گامی مهم در سنجش معتبر و احقاق حقوق افراد اقلیت است. در حال حاضر، عدالت در سنجش از جمله شواهد روایی مورد نیاز برای توجیه استفاده از آزمون برای یک هدف خاص است (سیرسی، ۲۰۱۶). انجمن روان‌شناسی اروپا برای سنجش کیفیت آزمون‌ها، کنش افتراقی سؤال (DIF) را به‌عنوان یکی از طرح‌های ممکن برای جمع‌آوری شواهد روایی سازه تعریف می‌کند (اورس و همکاران، ۲۰۱۳).

با توجه به موضوعات مطرح‌شده در حوزه سوگیری، بر اساس نتایج حاصل از سنجش روان‌شناختی تصمیماتی اتخاذ می‌گردد. هنگامی که آزمون دارای سوگیری اندازه‌گیری است، تصمیماتی که گرفته می‌شود از دقت، قوام و صحت کافی برخوردار نیستند. این تصمیمات نادرست که حاصل نقص در آزمون است، پیامدهای ناگواری از جمله در گزینش، ارتقاء، غربالگری و سایر حوزه‌ها دارد. لذا مؤسسات، سازمان‌ها، شرکت‌ها و وزارت‌خانه‌هایی که اقدام به برگزاری آزمون می‌کنند و نتیجه آزمون آنها در سرنوشت افراد، سازمان و حتی بخشی از کشور تأثیر می‌گذارد، ملزم هستند از عدالت و روایی سنجش خویش اطمینان حاصل نمایند.

شناسایی سؤالات سودار و حذف آنها از آزمون اقدامی سازنده است. مطالعات نشان داده است عوامل متعددی در کشف سؤالات سودار مؤثرند. مرور ادبیات پژوهش نشان می‌دهد ارزش پارامترهای سؤال در کیفیت سؤالات آزمون تأثیرگذار است. سؤالات آسان که حاصل ارزش پایین پارامتر دشواری است، با ایجاد اثر سقف و سؤالات دشوار با ایجاد اثر کف، پایایی و روایی آزمون را تحت‌الشعاع قرار داده و موجب سوگیری اندازه‌گیری می‌شوند (آلن و ین[12] ، ۱۳۹۵). پارامتر تمیز سؤال نیز به نوبه خود سوگیری اندازه‌گیری را تحت تأثیر قرار می‌دهد و با ایجاد تغییر در منحنی ویژگی سؤال در گروه‌های اقلیت و اکثریت، باعث بروز کنش افتراقی یکنواخت و غیریکنواخت می‌شود. هیدالگو و پینا[13] (۲۰۰۴) نشان دادند روش‌های متفاوت تشخیص کنش افتراقی سؤال با سطوح متفاوت پارامتر سؤال، عملکرد متفاوتی در تشخیص کنش افتراقی دارند. همچنین رابطه‌ی بین دشواری سؤال و کنش افتراقی نشان داد که بیشتر سؤالات دشوار گرایش به نشان دادن کنش افتراقی در گروه کانونی (معمولاً گروه‌های اقلیت) دارند. بنیتو[14] و همکاران (۲۰۱۸) اذعان

[1] Bauer
[2] Raju
[3] Reise
[4] Vandenberg
[5] American Psychological Association
[6] Byrne & Stewart
[7] Millsap
[8] Zumbo
[9] Camilli
[10] Hambleton
[11] Angoff
[12] Allen & Yen
[13] Hidalgo & Pina
[14] Benito

کرده‌اند هرگونه پارامتر آزمون یا سؤال از جمله ضرایب تمیز یا دشواری که بین دو یا چند گروه زیرجمعیت متفاوت است، اگر این تفاوت برای هر گروه تفسیرهای متفاوتی نیاز داشته باشد، ممکن است تهدیدی برای روایی آزمون باشد.

علاوه بر ضرایب دشواری و تمیز، حدس نیز در روایی و سوگیری آزمون تأثیر دارد. سانتلیس و ویلسون[1] (۲۰۱۲) ادعا کرده‌اند یکی از فرضیه‌هایی که نشان می‌دهد بین کنش افتراقی و ضریب دشواری آزمون ارتباط وجود دارد، می‌تواند به دلیل حدس تصادفی باشد. آزمودنی‌هایی که از روی حدس و گمان پاسخ صحیح را می‌یابند و در این کار خبره هستند، نمره بهتری کسب می‌کنند (سیف، ۱۳۹۷) که نتیجه آن باعث بیش‌برآورد توانایی و کاهش روایی و پایایی آزمون می‌گردد (تجالا و فیتریانی[2] ، ۲۰۱۷؛ مرتاض هجری، خباز مافی‌نژاد و جلیلی، ۱۳۹۳). حدس، ماتریس نمره‌ها را تغییر می‌دهد که در روایی و پایایی آزمون مؤثر است (مگنسون، ۱۳۷۰). حدس با به وجود آوردن یک خطای تصادفی باعث پایین آمدن روایی و پایایی آزمون، به دست آوردن نمرات غیرواقعی از افراد و در نتیجه عدم تشخیص درست سطح پیشرفت و تضییع حق کسانی که از حدس و گمان استفاده نمی‌کنند در مقابل کسانی که استفاده می‌کنند، می‌شود (گل‌محمدنژاد بهرامی، ۱۳۸۷).

با توجه به شرایط بررسی‌شده و نیاز به شناسایی سؤالات سودار، روش‌های تشخیص کنش افتراقی سؤال را به‌طور کلی می‌توان به روش‌های مبتنی بر مدل نظریه سؤال–پاسخ (IRT) و روش‌های مبتنی بر نمره آزمون (غیر IRT) تقسیم‌بندی کرد. علی‌رغم ریشه‌دار بودن موضوع و فعالیت‌های زیادی که در این زمینه صورت پذیرفته است، هنوز در مورد روش‌های مطلوب و بهینه تشخیص کنش افتراقی سؤال میان صاحب‌نظران اتفاق‌نظر وجود ندارد (میلسپ، ۲۰۱۱؛ راجو و همکاران، ۲۰۰۲؛ ریس[3] و همکاران، ۱۹۹۳؛ وندربرگ[4] ، ۲۰۰۲).

تحقیقات زیادی در مورد روش‌های تشخیص کنش افتراقی سؤال و تغییرناپذیری اندازه‌گیری انجام شده است که با پیشنهاداتی شامل استفاده از آزمون‌های نسبت احتمال[5] (تیشن[6] و همکاران، ۱۹۹۳)، آزمون والد[7](وود[8] و همکاران، ۲۰۱۳) و نمرات آزمون (شاخص‌های اصلاح یا آزمون‌های ضرب لاگرانژ[9] ؛ اورت[10] ، ۱۹۹۸) روبه‌رو هستیم. گرامی‌پور (۱۳۹۳) اذعان کرده روشی که عموماً مورد استفاده قرار می‌گیرد بر اساس آمار منتل–هنسل[11] است. داربینکوا و مارتینکوا[12] (۲۰۱۷) نیز اذعان داشته‌اند رگرسیون لجستیک و منتل–هنسل عمومی‌ترین و پرکاربردترین روش‌های تشخیص کنش افتراقی هستند، اما با این حال، هیچ‌یک از این روش‌ها اثر حدس را در نظر نمی‌گیرند. روش‌ها و مدل‌هایی که بتوانند عامل حدس را مفهوم‌سازی و کمّی‌سازی نموده و وارد معادلات کنند، بسیار محدودند. به‌طور کلی، از میان تمامی روش‌های معرفی‌شده برای مطالعه و کنترل اثر حدس در ردیابی کنش افتراقی سؤال، مدل‌های سه‌پارامتری نظریه پاسخ به سؤال و روش رگرسیون غیرخطی سعی در کمّی‌سازی مفهوم حدس و ارائه راهکارهایی جهت کنترل آن دارند.

مدل‌های نظریه سؤال–پاسخ در تشخیص کنش افتراقی دارای مشکلاتی مانند مفهوم متغیر پنهان، عدم برازش مناسب در مدل‌های سه‌پارامتری و نیاز به حجم نمونه بالا هستند. در بین روش‌های مبتنی بر نظریه سؤال–پاسخ، پرکاربردترین روش‌های دارای قابلیت احتساب اثر حدس، روش‌های لرد و راجو هستند که معایب خاص روش‌های نظریه سؤال–پاسخ را دارند (کیم و اوشیما[13] ، ۲۰۱۳).

روش رگرسیون غیرخطی، گسترشی از رگرسیون لجستیک دوپارامتری است که اجازه می‌دهد خط مجانب پایین متفاوت از صفر باشد و در ادبیات روانسنجی به‌عنوان مدل لجستیک سه‌پارامتری شناخته می‌شود (گلاس و فالکون[14] ، ۲۰۰۳). در روش رگرسیون غیرخطی، برآورد پارامترهای مدل به‌وسیله برآورد حداقل مجذورات غیرخطی تعیین می‌شود. برخلاف سایر روش‌های غیر نظریه سؤال–پاسخ، این روش عامل حدس را محاسبه می‌کند و به‌عنوان یک روش غیر نظریه سؤال–پاسخ به حساب می‌آید. مطالعات نشان می‌دهد این روش از ویژگی‌های مفیدی از جمله قدرت کافی و نرخ رد پایین برخوردار است (داربینکوا و مارتینکوا، ۲۰۱۷). رگرسیون غیرخطی برخلاف روش‌های نظریه سؤال–پاسخ، روشی نمره‌محور است و دانش آزمودنی‌ها به‌وسیله نمره استانداردشده کل آزمون مدل‌سازی شده است. به نظر راجو (۱۹۹۰)، مدل رگرسیون غیرخطی می‌تواند نماینده و جایگزینی برای مدل سه‌پارامتری نظریه سؤال–پاسخ برای تشخیص کنش افتراقی سؤال باشد. تشخیص اینکه آیا رگرسیون غیرخطی می‌تواند مدلی جایگزین جهت تشخیص کنش افتراقی سؤال به جای مدل سه‌پارامتری نظریه سؤال–پاسخ باشد، نیاز به بررسی تجربی دارد.

مطالعات و بررسی‌های انجام‌شده پیرامون انواع رویکردها و روش‌ها نشان می‌دهد که علی‌رغم سال‌ها مطالعه مدون که منجر به ارائه و معرفی روش‌های متعددی شده است، هنوز در استفاده از رویکرد و روش بهینه در تشخیص کنش افتراقی سؤال (DIF) ابهاماتی وجود دارد. تمامی این روش‌ها از کیفیت و مطلوبیت کافی برخوردار نیستند؛ چرا که برخی بسیار پیچیده و پرهزینه هستند و برخی دیگر با وجود سادگی، بی‌ثبات و غیردقیق‌اند. مهم‌تر اینکه اکثر آنها عامل حدس را که در تمامی آزمون‌های عینی وجود دارد، نادیده می‌گیرند و تنها روش‌های معدودی وجود دارند که این عامل را لحاظ

[1] Santelices &Wilson
[2] Tjalla & Fitriani
[3] Reise
[4] Vanderberg
[5] Likelihood Ratio
[6] Thissen
[7] Wald Test
[8] Woods
[9] Lagrange Multiplier
[10] Oort
[11] Mantel-Haenszel
[12] Drabinová & Martinková
[13] Kim & Oshima
[14] Glas & Falcón

می‌کنند. علاوه بر این، اثر همزمان پارامترهای سؤال بر میزان تشخیص صحیح کنش افتراقی توسط روش‌های مختلف روشن نیست. از بین روش‌های موجود، به روشی نیاز است که استانداردهای لازم را دارا باشد؛ بدین معنا که دقیق، ساده و قابل اجرا باشد و مهم‌تر از همه بتواند پارامترهای دشواری و تمیز را در سوگیری مورد توجه قرار داده و نسبت به اثر حدس حساس باشد. بنابراین، ضروری است که عامل حدس در سوگیری اندازه‌گیری مفهوم‌سازی و کمّی‌سازی شده و اثر آن در سنجش مورد توجه قرار گیرد. اگر از این مفهوم‌سازی و اندازه‌گیری غفلت شود، اندازه‌گیری با مشکلات عدیده‌ای مواجه خواهد شد؛ چنان‌که احتمال دارد معنای نمرات به اشتباه تفسیر شده و ضمن کاهش دقت نمرات، عواقب پیش‌بینی نشده‌ای در پی داشته باشد.

هرچند روش‌های متعددی معرفی شده‌اند و اگرچه به لحاظ نظری رویکرد رگرسیونی نسبت به رویکردهای مبتنی بر نظریه سؤال–پاسخ یا روش‌های پرکاربرد کلاسیک برتری داشته و قابلیت مقابله با چالش‌های این روش‌ها را دارد، با این همه، شواهد تجربی اندکی در خصوص دامنه کاربرد، عوامل و شرایط تأثیرگذار بر آن وجود دارد و پژوهش دقیق، مشخص و مدونی در خصوص کارایی این روش انجام نشده است. بنابراین، لازم است مطالعه‌ای جامع و کامل طراحی شود که با در نظر گرفتن شرایط و مؤلفه‌های تأثیرگذار به تفکیک، موضوع را مورد بررسی قرار دهد و نتایج آن به‌عنوان یک راهنمای عملی در اختیار آزمون‌سازان، گزینش‌گران و سایر متخصصان قرار گیرد. لذا هدف از پژوهش حاضر، شناسایی رویکرد بهینه سه‌پارامتری تشخیص کنش افتراقی سؤال در شرایطی است که امکان حدس وجود دارد و مؤلفه‌های دیگری مانند ضرایب تمیز و دشواری سؤالات نیز تأثیرگذار هستند.

روش پژوهش

پژوهش حاضر به لحاظ هدف، کاربردی؛ از نظر روش، کمّی و در زمره مطالعات روانسنجی قرار می‌گیرد. انجام این مطالعه مستلزم اعمال دستکاری در سطوح مختلف عوامل مورد بررسی و مطالعه نتایج حاصل از آن بود. جهت دستیابی به این هدف، از روش شبیه‌سازی استفاده شد. به دلیل پرهزینه و غیرعملی بودن گردآوری داده‌های واقعی و وجود نقص در این داده‌ها، به‌ویژه زمانی که داده‌های گمشده وجود دارد و الگوی آنها تصادفی نیست، پژوهشگران و روان‌سنجان شبیه‌سازی داده‌ها را به جای گردآوری داده‌های واقعی ترجیح می‌دهند (بولت و سانبول[1] ، ۲۰۱۷).

در پژوهش حاضر از سه متغیر مستقل و یک متغیر وابسته استفاده شد. متغیرهای مستقل عبارت بودند از: روش‌های سه‌پارامتری تشخیص کنش افتراقی سؤال (لرد، راجوی سه‌پارامتری و رگرسیون غیرخطی)، سطوح پارامترها (کم، متوسط و زیاد) و پارامترهای سؤال (دشواری، تمیز و حدس). متغیر وابسته نیز شامل میزان تشخیص صحیح کنش افتراقی سؤال بود. روش اجرا بدین صورت بود که ابتدا بر اساس اهداف پژوهش، داده‌ها با استفاده از نرم‌افزار WINGEN-3 تولید شدند؛ سپس با بهره‌گیری از نرم‌افزار R-3.6 تحلیل‌های تشخیص کنش افتراقی انجام گرفت. در نهایت، تحلیل واریانس با استفاده از نرم‌افزار SPSS-26 بر روی اطلاعات حاصل از میزان تشخیص صحیح کنش افتراقی مربوط به هر یک از روش‌ها صورت پذیرفت.

الف) شبیه‌سازی: در مرحله شبیه‌سازی، ۲۰ درصد از سؤالات با شدت اثر ۰/۵ مورد دستکاری آزمایشی قرار گرفته و به‌عنوان سؤالات دارای کنش افتراقی وارد مطالعه شدند. جهت دستیابی به نتایج پایدار، هر نوع از داده‌ها با توجه به پارامترهای سؤال و سطوح سه‌گانه، در ۱۰ مجموعه تولید گردید که در مجموع ۹۰ مجموعه داده شبیه‌سازی شد. توزیع توانایی در تمامی داده‌های تولیدشده از نوع نرمال استاندارد و پاسخ‌های تولیدی، دوارزشی و مبتنی بر مدل سه‌پارامتری بود. همچنین، پارامترهای سؤال از توزیع یکنواخت با دامنه پیوسته پیروی می‌کردند.

جدول ۱

سطوح و بازه‌های مورد استفاده برای پارامترهای شبیه‌سازی سؤالات

متغیر مستقل	طیف کم	طیف متوسط	طیف زیاد	منبع
دشواری	۳– تا ۱–	۱– تا ۱	۱ تا ۳	هیدالگو و پینا (۲۰۰۴)، نارایانون و سوامیناتان[2] (۱۹۹۶)، و بیکر (۲۰۰۴)
تمیز	۰/۲۵ تا ۰/۷۵	۰/۷۵ تا ۱/۲۵	۱/۲۵ تا ۱/۷۵	
حدس	۰ تا ۰/۱	۰/۱ تا ۰/۲	۰/۲ تا ۰/۳	گلس و فالکون (۲۰۰۳)، بیکر (۲۰۰۴)، داربینکوا و مارتینکوا (۲۰۱۷)

بر اساس اطلاعات مندرج در جدول فوق و با تکیه بر پیشینه پژوهشی، برای هر یک از پارامترها سه سطح جهت شبیه‌سازی تعیین شده است. شایان ذکر است مطالعاتی که تأثیر عامل حدس بر سوگیری را بررسی کرده باشند، بسیار محدود هستند؛ که این امر احتمالاً ناشی از محدودیت‌های جدی روش‌های آماری در مفهوم‌سازی و مدل‌سازی اثر حدس در تشخیص کنش افتراقی است. در حالی که در اکثر مطالعات شبیه‌سازی پیشین برای پارامتر حدس صرفاً از مقدار ثابت ۰/۲ استفاده می‌شد، در پژوهش حاضر به منظور بررسی دقیق‌تر، سه سطح کم، متوسط و زیاد برای این پارامتر در نظر گرفته شده است.

ب) تحلیل‌های کنش افتراقی: جهت انجام تحلیل‌های کنش افتراقی، از نسخه ۳/۶ نرم‌افزار R استفاده شد. روش رگرسیون غیرخطی با به‌کارگیری تابع جدید difNLR در بسته difNLR (داربینکوا و همکاران، ۲۰۱۶) اجرا گردید که برای برآورد حداقل مجذورات غیرخطی از تابع nls موجود در بسته stats بهره می‌برد (دنیس و همکاران، ۱۹۸۱؛ ریتز و استریبیگ، ۲۰۰۸). مدل سه‌پارامتری سؤال–پاسخ برای تمامی داده‌ها با استفاده از تابع

[1] Bulut & Sunmul

[2] Narayanon & Swaminathan

itemParEst برازش داده شد و پارامترهای حدس برآورد گردیدند. سپس، ضرایب برآوردشده مجدداً مقیاس‌بندی شده و آماره‌های لرد و راجو با استفاده از توابع difLord و difRaju محاسبه شدند. مقادیر P برای آماره لرد بر اساس توزیع خی‌دو (χ^2) با ۲ درجه آزادی و برای آماره راجو بر اساس توزیع نرمال استاندارد تعیین گردید. در نهایت، برای هر مجموعه از داده‌ها ۳ آزمون اجرا شد که مجموعاً منجر به انجام ۲۷۰ آزمون تشخیص کنش افتراقی گردید.

ج) تحلیل واریانس: نتایج حاصل از تحلیل‌های کنش افتراقی در چهار حالت شامل مثبت صحیح، مثبت کاذب، منفی صحیح و منفی کاذب طبقه‌بندی شدند. این حالات در نهایت در دو دسته کلی «تشخیص صحیح» (مجموع مثبت صحیح و منفی صحیح) و «تشخیص نادرست» (مجموع مثبت کاذب و منفی کاذب) قرار گرفتند. منظور از تشخیص صحیح، شناسایی کنش افتراقی در سؤالاتی است که دارای این ویژگی بودند (۲۰ درصد سؤالات دستکاری‌شده) و عدم شناسایی آن در سؤالاتی است که فاقد این ویژگی بودند (۸۰ درصد سؤالات دستکاری‌نشده). میزان تشخیص صحیح به‌عنوان متغیر وابسته وارد مدل تحلیل واریانس شد. همچنین، پیش‌فرض‌های آزمون تحلیل واریانس از جمله نرمال بودن توزیع داده‌ها و همگنی واریانس‌ها بررسی و مورد تأیید قرار گرفتند.

یافته‌ها

پس از تولید نظام‌مند داده‌های شبیه‌سازی‌شده و اجرای فرآیند تشخیص کنش افتراقی سؤال (DIF) با استفاده از روش‌های تعیین‌شده، مجموعه‌ای جامع از نتایج کمی فراهم آمد. مرحله حاضر متوجه استخراج معنای علمی از این داده‌ها و پاسخ به پرسش‌های اصلی پژوهش است. هدف از تحلیل‌های آماری پیش‌رو، نه توصیف فرآیند اجرا، که ارزیابی مقایسه‌ای دقت و کارایی روش‌های مختلف تشخیص DIF تحت شرایط کنترل‌شده و شناسایی الگوهای تأثیر عوامل مداخله‌گر بود. تحلیل‌ها در دو سطح کلان و خرد طراحی شدند. در سطح کلان، از تحلیل واریانس سه‌راهه برای بررسی همزمان اثرات اصلی و تعاملی سه عامل بنیادی «یعنی نوع روش آماری، ماهیت پارامتر سؤال و سطح مقادیر آن» بر نرخ تشخیص صحیح استفاده شد. این تحلیل به‌طور مستقیم به پرسش اصلی مطالعه درباره برتری روش‌ها و شرایط بهینه عملکرد آنها پاسخ می‌دهد. محققان در سطح خرد و با هدف واکاوی دقیق‌تر تعاملات معنادار شناسایی‌شده، به سراغ تحلیل‌های اثرات ساده و مقایسه‌های زوجی رفتند. این رویکرد امکان داد تا مشخص شود تفاوت‌های مشاهده‌شده بین روش‌ها یا بین سطوح مختلف یک پارامتر، دقیقاً در کدام شرایط (مثلاً در سطح «کم» یک پارامتر خاص) رخ می‌دهد و جهت این تفاوت‌ها چگونه است. خروجی این تحلیل‌ها فراتر از ارائه جداول آماره‌های عددی است. آنها امکان تفسیر عملی یافته‌ها را فراهم می‌کنند و مشخص می‌سازند که کدام روش، تحت کدام ترکیب از ویژگی‌های سؤال، قابل‌اعتمادترین ابزار برای پژوهشگران و متخصصان سنجش به‌شمار می‌رود. در ادامه، نتایج حاصل از این ارزیابی‌های آماری نظام‌مند به‌تفصیل ارائه می‌گردد.

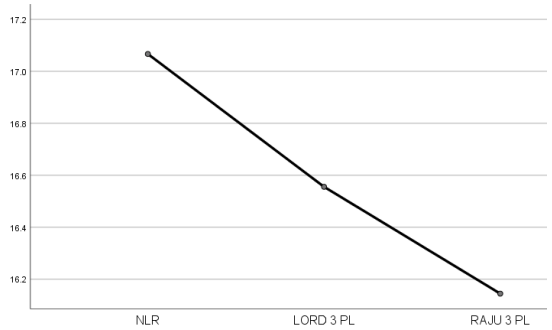
جدول ۲

نتیجه تحلیل واریانس سه راهه (پارامترهای سؤال، طیف پارامترها و نوع آزمون)

منبع	SS	df	MS	F	a	η^2
عرض از مبدا	۷۴۳۰۱	۱	۷۴۳۰۱	۲۲۸۸۰	p <۰/۰۵	۰/۹۸۹
پارامترها	۲۶۰	۲	۱۳۰	۴۰	p <۰/۰۵	۰/۲۴۸
طیف (کم، متوسط و زیاد)	۱۲۸	۲	۶۴	۱۹	p <۰/۰۵	۰/۱۴۰
نوع آزمون	۳۸	۲	۱۹	۶	p <۰/۰۵	۰/۰۴۶
تعامل پارامترها و طیف	۱۴۹	۴	۳۷	۱۱	p <۰/۰۵	۰/۱۵۹
تعامل پارامترها و آزمون	۴	۴	۰/۹۶	۰/۳۰	p >۰/۰۵	۰/۰۰۵
تعامل طیف و آزمون	۱۴	۴	۳	۲	p >۰/۰۵	۰/۰۱۷
تعامل سه راهه	۴۷	۸	۶	۲	p >۰/۰۵	۰/۰۵۶
خطا	۷۸۹	۲۴۳	۳			
کل	۷۵۷۳۱	۲۷۰				

داده‌های حاصل از تشخیص صحیح کنش افتراقی با استفاده از آزمون تحلیل واریانس سه‌راهه بین‌گروهی مورد تجزیه‌وتحلیل قرار گرفت. در این تحلیل، متغیر وابسته «میزان تشخیص صحیح» و متغیرهای مستقل شامل «پارامترهای سؤال» (در سه سطح: دشواری، تمیز و حدس)، «طیف پارامترها» (در سه سطح: کم، متوسط و زیاد) و «آزمون‌های سه‌پارامتری» (در سه سطح: راجو، لرد و رگرسیون غیرخطی) بودند. نتایج تحلیل واریانس سه‌راهه نشان داد که هر سه اثر اصلی مورد بررسی از نظر آماری معنادار هستند. پارامترهای سؤال با اندازه اثر (η^2=۰/۲۴۸)، طیف پارامترها با اندازه اثر (η^2=۰/۱۴۰) و نوع آزمون با اندازه اثر (η^2=۰/۰۴۶) همگی در سطح p<۰/۰۵ معنادار بودند. این یافته‌ها حاکی از آن است که تغییرات در هر یک از این متغیرهای مستقل به‌تنهایی بر میزان تشخیص صحیح کنش افتراقی تأثیر دارد. در میان این سه، پارامترهای سؤال بیشترین سهم را در تبیین واریانس متغیر وابسته داشت و پس از آن به‌ترتیب طیف پارامترها و نوع آزمون قرار گرفتند. در میان اثرهای تعاملی دوطرفه موجود در مدل، تنها تعامل میان «پارامترهای سؤال»

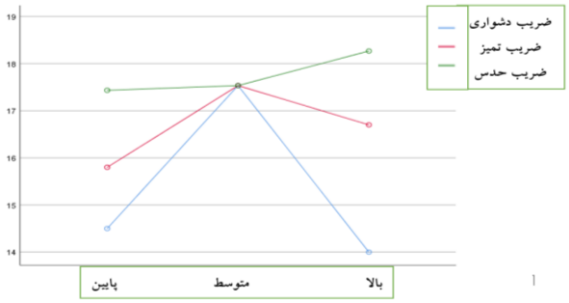
و «طیف پارامترها» از نظر آماری معنادار بود و تعامل سه‌طرفه معنادار نشد. با توجه به عدم معناداری تعاملات ۲ و ۳ طرفه با حضور عامل آزمون، آزمون تعقیبی توکی برای اثر اصلی متغیر آزمون اجرا و نتیجه نشان داد آزمون رگرسیون غیرخطی عملکرد بهتر و معناداری از آزمون راجو سه پارامتری دارد.



نمودار ۱

میزان تشخیص صحیح DIF بر اساس نوع آزمون (اثر اصلی آزمون)

طبق نمودار ۱ آزمون رگرسیون غیرخطی بیشترین توان تشخیصی و آزمون سه‌پارامتری راجو کمترین توان را در تشخیص صحیح DIF دارند.



نمودار ۲

میانگین تشخیص صحیح DIF بر اساس پارامترهای سؤال (دشواری، تمیز و حدس) و سطوح طیف (کم، متوسط و زیاد)

با توجه به معناداری تعامل دوطرفه بین پارامترهای سؤال و سطوح طیف، نمودار نشان می‌دهد که میزان تشخیص صحیح DIF هنگامی که پارامترهای دشواری و تمیز در سطح متوسط قرار دارند، بیشتر از مواقعی است که این پارامترها در سطوح بالا یا پایین باشند. همچنین پارامتر حدس نشان می‌دهد که با افزایش مقدار ضریب حدس، روش‌های سه‌پارامتری عملکرد بهتری از خود نشان می‌دهند. از آنجا که تعامل دوطرفه بین پارامترهای سؤال و سطوح طیف معنادار است، تفسیر مستقیم اثرات اصلی امکان‌پذیر نیست؛ بنابراین باید اثرات ساده مورد بررسی قرار گیرند که در ادامه آمده است.

الف: اثرات ساده پارامترها روی طیف‌ها

جدول ۳

نتیجه تحلیل واریانس یک راهه (اثرات ساده پارامترها روی طیف‌ها)

مرجع	منبع	SS	df	MS	F	P	η^2	آزمون تعقیبی
طیف کم	عرض از مبدأ	۲۲۷۸۴	۱	۲۲۷۸۴	۱۰۲۳۵	p <۰/۰۵	۰/۹۹۲	حدس > دشواری حدس> تمیز تمیز> دشواری
	پارامترها	۱۲۹	۲	۶۴	۲۹	p <۰/۰۵	۰/۴۰۱	
	خطا	۱۹۳	۸۷	۲				
	کل	۲۳۱۰۸	۹۰			R^2 = ۰/۴۰۱		
طیف متوسط	عرض از مبدأ	۲۷۶۶۷۰	۱	۲۷۶۶۷۰	۱۱۸۹۲۰	p <۰/۰۵	۰/۹۹۳	
	پارامترها	۰/۰۰	۲	۰/۰۰	۰/۰۰	p <۰/۰۵	۰/۰۰۰	
	خطا	۲۰۲۰	۸۷	۲۰				
	کل	۲۷۸۷۰۰	۹۰			R^2=۰/۰۰		
طیف زیاد	عرض از مبدأ	۲۳۹۷۷۰	۱	۲۳۹۷۷	۴۲۰۴	p <۰/۰۵	۰/۹۸۰	حدس > دشواری تمیز> دشواری
	پارامترها	۲۷۹۰	۲	۱۳۹۰	۲۴	p <۰/۰۵	۰/۳۶۰	
	خطا	۴۹۶۰	۸۷	۵				
	کل	۲۴۷۵۳۰	۹۰			R^2=۰/۳۶		

نتایج تحلیل واریانس یک‌طرفه اثرات ساده پارامترها بر حسب سطوح طیف نشان داد عملکرد پارامترها در سطح کم (F(۲، ۸۷)=۲۹، P<۰/۰۵، η^2=۰/۴۰۱) و طیف زیاد (F(۲، ۸۷)=۲۴، P<۰/۰۵، η^2=۰/۳۶۰) معنادار است. آزمون تعقیبی توکی نشان داد هنگامی که سطح طیف در تمام پارامترها کم است، پارامتر حدس نسبت به پارامترهای تمیز و دشواری تأثیر مثبت و معناداری بر میزان تشخیص صحیح DIF دارد و همچنین پارامتر تمیز نسبت به دشواری برتری معناداری دارد. وقتی سطح طیف در حالت متوسط قرار دارد، تفاوت معناداری بین تأثیر پارامترها مشاهده نشد. در سطح بالا نیز پارامتر حدس نسبت به دشواری تأثیر مثبت و معناداری نشان می‌دهد و پارامتر تمیز در مقایسه با دشواری برتری معناداری دارد.

ب: اثرات ساده طیف‌ها روی پارامترها

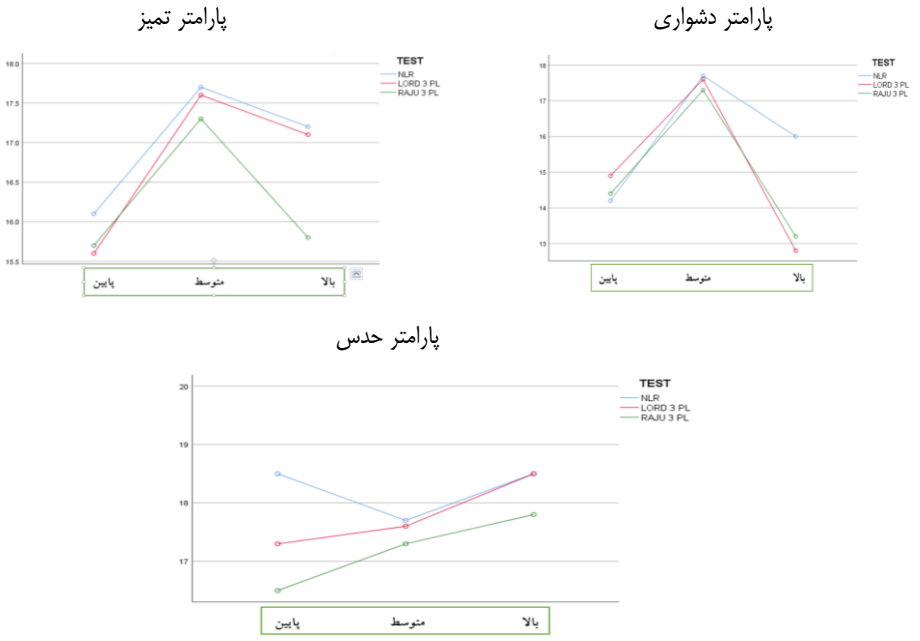
جدول ۴

نتیجه تحلیل واریانس یک راهه (اثرات ساده طیف‌ها روی پارامترها)

مرجع	منبع	SS	df	MS	F	P	η^2	آزمون تعقیبی
ضریب دشواری	عرض از مبدأ	۲۱۱۹۰	۱	۲۱۱۹۰	۳۷۳۹	p <۰/۰۵	۰/۹۷۷	متوسط > کم
	طیف	۲۱۹	۲	۱۰۹	۱۹	p <۰/۰۵	۰/۳۰۸	متوسط > زیاد
	خطا	۴۹۲	۸۷	۵				
	کل	۲۱۹۰۳	۹۰			R^2=۰/۳۰۸		
ضریب تمیز	عرض از مبدأ	۲۵۰۳۳	۱	۲۵۰۳۳	۱۱۸۰۰	p <۰/۰۵	۰/۹۹۳	متوسط > کم
	طیف	۴۵	۲	۲۲	۱۰	p <۰/۰۵	۰/۱۹۶	بالا > کم
	خطا	۱۸۴	۸۷	۲				
	کل	۲۵۲۶۳	۹۰			R^2= ۰/۱۹۶		
ضریب حدس	عرض از مبدأ	۲۸۳۳۷	۱	۲۸۳۳۷	۱۱۴۸۲	p <۰/۰۵	۰/۹۹۲	
	طیف	۱۲	۲	۶	۲	p >۰/۰۵	۰/۰۵۵	
	خطا	۲۱۴	۸۷	۲				
	کل	۲۸۵۶۵	۹۰			R^2= ۰/۰۵۵		

نتایج تحلیل واریانس یک‌راهه اثرات ساده طیف‌ها روی پارامترها نشان می‌دهد عملکرد طیف‌ها با وجود پارامتر دشواری (F(۲، ۸۷)=۱۹، P<۰/۰۵، η^2=۰/۳۰۸) و پارامتر تمیز (F(۲، ۸۷)=۱۰، P<۰/۰۵، η^2=۰/۱۹۶) معنادارند.

نتایج مقایسه‌ای آزمون تعقیبی توکی نشان داد وقتی پارامتر دشواری در سطح متوسط قرار دارد، میزان تشخیص صحیح کنش افتراقی توسط آزمون‌ها بیشتر از زمانی است که دشواری در سطح کم یا زیاد باشد. همچنین وقتی پارامتر تمیز در سطح متوسط قرار دارد، میزان تشخیص صحیح کنش افتراقی نسبت به حالت کم بیشتر است و این برتری بین تمیز بالا و تمیز پایین نیز مشاهده شد. پارامتر حدس در تمام سطوح طیف عملکرد مشابهی نشان داد.



نمودار ۳

میزان تشخیص صحیح DIF توسط آزمون‌ها بر حسب پارامترهای سؤال (دشواری، تمیز و حدس) در سطوح مختلف طیف (کم، متوسط و زیاد)

نمودارهای ترکیبی نشان می‌دهد آزمون رگرسیون غیرخطی در تمامی سطوح پارامترها (به‌جز سطح پایینِ دشواری) برتری دارد؛ آزمون راجو تقریباً ضعیف‌ترین عملکرد و آزمون لرد عملکردی میان‌رده نشان داد. همچنین در تحلیل ترکیبی آزمون‌ها، طیف‌ها و پارامترها، برای پارامترهای دشواری و تمیز تمام آزمون‌ها در سطح متوسط عملکرد بهتری نسبت به سطوح پایین یا بالا از خود نشان دادند. در پارامتر حدس، راجو و لرد با افزایش مقدار ضریب حدس میزان تشخیص صحیح را افزایش می‌دهند، اما رگرسیون غیرخطی در مقدار متوسط ضریب حدس افتی در تشخیص صحیح نشان می‌دهد؛ هرچند آزمون تحلیل واریانس اثر پارامتر حدس را معنادار نشان نداد.

بحث و نتیجه‌گیری

هدف این پژوهش شناسایی رویکرد بهینه در تشخیص کنش افتراقی در شرایطی بود که امکان حدس وجود دارد و پارامترهایی نظیر دشواری و تمیز سؤالات نیز مؤثر باشند. چنین پژوهش‌هایی می‌تواند به ارتقای کارایی آزمون‌ها کمک کند (باقری‌خواه و همکاران، ۱۳۹۰)، شناسایی منابع تهدیدکننده‌ی تغییرناپذیری اندازه‌گیری را تسهیل نماید (بائر و همکاران، ۲۰۲۰)، در تضمین همگنی اندازه‌گیری مؤثر باشد (راجو و همکاران، ۲۰۰۲؛ ریس و همکاران، ۱۹۹۳؛ وندربرگ، ۲۰۰۲) و پایه‌ای برای اتخاذ تصمیمات منطقی در فرایندهای ارزیابی فراهم آورد (زومبو، ۱۹۹۹).

به‌طور کلی یافته‌های پژوهش نشان داد میزان تشخیص صحیح DIF توسط روش‌های سه‌پارامتری تحت تأثیر پارامترهای دشواری، تمیز و حدس قرار دارد و این روش‌ها نسبت به تغییرات مقادیر پارامترهای سؤال حساسیت معنادار و قابل‌توجهی نشان می‌دهند. با مداخلات آزمایشی در مقادیر پارامترها مشخص شد هنگامی که پارامترهای دشواری و تمیز در میانه‌ی منحنی ویژگی سؤال قرار گیرند، توان شناسایی صحیح DIF در همه‌ی روش‌های سه‌پارامتری به حداکثر می‌رسد؛ در حالی که تغییر مقادیر دشواری یا تمیز به سمت چپ یا راست منحنی (کاهش یا افزایش مقادیر این ضرایب) منجر به کاهش توان تشخیص صحیح در همه‌ی روش‌ها می‌شود. این کاهش در سطح بالای ضریب تمیز معنادار نبود. بنابراین زمانی که سؤالات بسیار سخت یا بسیار آسان باشند، توان شناسایی صحیح نسبت به حالتی که سؤالات در حد متوسط دشواری قرار دارند به‌طور محسوس کاهش می‌یابد. این نتیجه با بخشی از یافته‌های هیدالگو و پینا (۲۰۰۴) و بنیتو و همکاران (۲۰۱۸) هم‌راستا است؛ از این رو بهترین جایگاه طراحی سؤالات آزمون حالتی است که منحنی ویژگی سؤال در میانه‌ی دامنه و دارای شیب متوسط باشد.

در بازه‌های مختلف ضریب حدس، عملکرد روش‌های تشخیص کنش افتراقی تا حد زیادی مشابه بود؛ ازاین‌رو ادعای سانتلیس و ویلسون (۲۰۱۲) مبنی بر ارتباط میان کنش افتراقی و ضریب دشواری آزمون به‌واسطه حدس تصادفی که بر رابطه‌ی تعاملی میان پارامترهای سؤال تمرکز دارد با یافته‌های پژوهش حاضر همسویی کامل ندارد و نیازمند بررسی‌های تجربی بیشتر است. یکی از دلایل احتمالی این ناهمسویی آن است که در مطالعات گذشته برای مفهوم‌سازی و کمی‌سازی اثر حدس از روش‌هایی استفاده شده که پارامتر حدس را در مدل لحاظ نمی‌کردند؛ در حالی که در این تحقیق، روش‌های سه‌پارامتری توانستند اثرات هر یک از پارامترهای سؤال را به‌صورت مجزا و ترکیبی آشکار سازند. با توجه به تأکید مکرر پژوهشگران بر تأثیر حدس بر روایی و خطای برآورد توانایی (سیف، ۱۳۹۷؛ تجالا و فیتریانی، ۲۰۱۷؛ مرتاض هجری و همکاران، ۱۳۹۳؛ گل‌محمدنژاد بهرامی، ۱۳۸۷)، پیشنهاد می‌شود بررسی‌های تجربی مبتنی بر داده‌های واقعی و با بهره‌گیری از مدل‌های شامل پارامتر حدس برای روشن‌تر شدن رابطه‌ی بین حدس، دشواری و کنش افتراقی انجام شود.

اگرچه ادبیات پژوهشی نشان می‌دهد از نظر رویکرد، روش‌شناسی و سایر عوامل مؤثر در تشخیص کنش افتراقی، تحلیل‌گران با وضعیت‌های تا حدودی مبهمی روبه‌رو هستند، یافته‌های پژوهش حاضر نشان می‌دهد در شرایط مختلف می‌توان روش‌های با بیشترین کارایی را مشخص و راهنمایی‌های کاربردی ارائه داد. به‌طور کلی انتخاب یک روش واحد برای تشخیص صحیح DIF تابعی از چند ویژگی داده‌ها و شرایط مطالعه است؛ بنابراین محقق باید با توجه به الگوی داده‌ها، حضور یا عدم حضور پدیده‌هایی چون حدس و نیز ویژگی‌های آیتم‌ها نظیر دشواری و تمیز، روش مناسب را برگزیده و در صورت امکان حساسیت‌سنجی بین روش‌ها انجام دهد تا از انتخاب رویکردی که بهترین نتیجه را در زمینه‌ی مورد بررسی فراهم می‌آورد اطمینان حاصل شود.

اگرچه این پژوهش با طراحی نظام‌مند شبیه‌سازی و کنترل دقیق متغیرها، بینشی ارزشمند از رفتار روش‌های سه‌پارامتری تشخیص DIF ارائه کرد، اما با محدودیت‌هایی نیز همراه بود. نخست، نتایج این مطالعه مبتنی بر داده‌های شبیه‌سازی‌شده تحت شرایط ایده‌آل (مانند توزیع نرمال توانایی و برازش کامل مدل) است. در دنیای واقعی، داده‌ها اغلب با مواردی مانند داده‌های پرت، نقص برازش مدل یا الگوهای پاسخ پیچیده‌تر مواجه هستند که می‌توانند عملکرد روش‌ها را تحت تأثیر قرار دهند. دوم، این پژوهش تنها سه روش پرکاربرد سه‌پارامتری را بررسی کرد. مقایسه عملکرد روش رگرسیون غیرخطی با روش‌های جدیدتر یا رویکردهای مبتنی بر یادگیری ماشین می‌تواند افق‌های تازه‌ای بگشاید.

بر این اساس، پیشنهاد می‌شود پژوهش‌های آتی با به‌کارگیری داده‌های واقعی از حوزه‌های مختلف (مانند آزمون‌های تحصیلی، پرسشنامه‌های بالینی)، صحت یافته‌های این مطالعه را در شرایط غیرایده‌آل بیازمایند. همچنین، گسترش مقایسه‌ها به روش‌های دیگر (همچون روش‌های مبتنی بر بیز، یا رویکردهای نیمه‌پارامتری) و بررسی تأثیر متغیرهای مداخله‌گر جدید (مانند طول آزمون، ناهمگنی واریانس خطا، یا حضور DIF غیریکنواخت) می‌تواند نقشه راه کامل‌تری برای انتخاب روش بهینه ترسیم کند. در نهایت، تدوین راهنمای عملیاتی یا درخت تصمیم‌گیری بر اساس یافته‌های این پژوهش و مطالعات مشابه، می‌تواند متخصصان سنجش و روانسنجی را در انتخاب روش تشخیص DIF متناسب با ویژگی‌های داده‌ها و اهداف پژوهش یاری رساند.

سپاس‌گزاری

محققان بر خود لازم می‌دانند از همه اساتید که با راهنمایی و ارائه اطلاعات ارزشمند خویش باعث تولید اثر علمی حاضر شدند، نهایت تقدیر و قدردانی را به عمل آورند.

ملاحظات اخلاقی

با توجه به استفاده از داده‌های شبیه‌سازی شده در پژوهش حاضر مشارکت‌کننده انسانی یا حیوانی وجود نداشت و موضوعات اخلاقی مطرح نشد.

حامی مالی

تحقیق حاضر فاقد حامی مالی است.

مشارکت نویسندگان

مسعود گرامی‌پور در انتخاب موضوع و چارچوب نظری، محمدرضا فلسفی‌نژاد در روش‌شناسی پژوهش و احمد رضائی در اجرا و نگارش اولیه مشارکت کرده‌اند. نسخه نهایی مقاله توسط هر سه نویسنده بررسی و تأیید شده است.

تعارض منافع

نویسندگان اعلام می‌کنند هیچ گونه تعارض منافعی نداشته‌اند.

منابع

آلن، م. و ین، و. (۱۳۹۵). *مقدمه‌ای بر نظریه‌های اندازه‌گیری (روانسنجی)*، (علی دلاور، مترجم). تهران: سمت.

باقری‌خواه، ز.، عارفی، م. و جمالی، ا. (۱۳۹۰). وضعیت پذیرش دانشجو در آموزش عالی ایران از دیدگاه دانشجویان دانشگاه‌های دولتی شهر تهران، اعضای هیأت‌علمی سازمان سنجش و مسئولان ذی‌ربط آموزش عالی. *اندازه‌گیری تربیتی*، ۲(۶)، ۱–۳۱.

سیف، ع. ا. (۱۳۹۷). *اندازه‌گیری، سنجش و ارزشیابی آموزشی*. تهران: نشر دوران.

گرامی‌پور، م. (۱۳۹۳). *مبانی نظری و کاربرد نظریه‌های اندازه‌گیری در علوم رفتاری*. تهران: تمدن علمی.

گل‌محمدنژاد بهرامی، غ. (۱۳۸۷). نقد اصول اساسی کاربرد فرمول حذف عامل شانس (با استفاده از نمره‌ی منفی) در آزمون‌ها. *نامه‌ی آموزش عالی*، ۱(۴)، ۵۱–۶۴.

مرتاض هجری، س.، خباز مافی‌نژاد، م. و جلیلی، م. (۱۳۹۳). پاسخ حدسی به سؤالات چندگزینه‌ای: چالش‌ها و راه‌کارها. *مجله ایرانی آموزش در علوم پزشکی*، ۱۴(۷)، ۵۹۴–۶۰۴.

مگنسون، د. (۱۳۷۰). *مبانی نظری آزمون‌های روانی* (محمدنقی براهنی، مترجم). تهران: دانشگاه تهران.

همبلتون، ر. ک.، سوامیناتان، ا. و راجرز، ه. ج. (۱۹۸۹/۱۳۸۹). *مبانی نظریه پرسش–پاسخ* (محمدرضا فلسفی‌نژاد، مترجم). تهران: دانشگاه علامه طباطبایی.

References

Allen, M., & Yen, W. (2016). *Introduction to Measurement Theories (Psychometrics)* (A. Delavar, Trans.). Tehran: SAMT. [Persian]

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–24). Lawrence Erlbaum Associates.

Bagherikhah, Z. , Arefi, M. and jamali, E. (2011). Situation analysis of student admission in Iranian higher education system from students, NOET's academicians and related educational officials' point of view. *Quarterly of Educational Measurement*, *2*(6), 1-31. [Persian]

Baker, F. B., & Kim, S. H. (Eds.). (2004). *Item response theory: Parameter estimation techniques*. CRC Press.

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods, 22*(3), 507–524.

Bauer, D. J., Belzak, W. C., & Cole, V. T. (2020). Simplifying the assessment of measurement invariance over multiple background variables: Using regularized moderated nonlinear

factor analysis to detect differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(1), 43–55.

Bulut, O., & Sünbül, Ö. (2017). Monte Carlo simulation studies in item response theory with the R programming language. *Journal of Measurement and Evaluation in Education and Psychology, 8*(3), 266–287.

Byrne, B. M., & Stewart, S. M. (2006). Teacher's corner: The MACS approach to testing for multigroup invariance of a second-order structure: A walk through the process. *Structural Equation Modeling, 13*(2), 287–321.

Camilli, G. (2006). Test fairness. *Educational Measurement, 4*, 221–256.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Vol. 4). Sage.

Dennis, J. E., Jr., Gay, D. M., & Walsh, R. E. (1981). An adaptive nonlinear least-squares algorithm. *ACM Transactions on Mathematical Software, 7*(3), 348–368.

Drabinová, A., & Martinková, P. (2017). Detection of differential item functioning with nonlinear regression: A non-IRT approach accounting for guessing. *Journal of Educational Measurement, 54*(4), 498–517.

Evers, A., Muñiz, J., Hagemeister, C., Høstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, D. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema, 25*(3), 283–291.

Geramipour, M. (2014). *Theoretical Foundations and Application of Measurement Theories in Behavioral Sciences*. Tehran: Elm-e-Tamadon. [Persian]

Glas, C. A., & Falcón, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27*(2), 87–106.

Golmohammad Nazhad Bahrami,G. (2009). Criticizing the Fundamental Principles of the Application of Chance Factor Elimination Formula in Test. *Higher Education Letter*, *1*(4), 51-64. [Persian]

Gómez-Benito, J., Sireci, S., Padilla, J. L., Hidalgo, M. D., & Benítez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema, 30*(1), 104–109.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (2010/1989). *Fundamentals of Item Response Theory* (M. R. Falsafi-Nejad, Trans.). Tehran: Allameh Tabataba'i University. [Persian]

Hidalgo, M. D., & López-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*(6), 903–915.

Kim, J., & Oshima, T. C. (2013). Effect of multiple testing adjustment in differential item functioning detection. *Educational and Psychological Measurement, 73*(3), 458–470.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*(6), 935–953.

Magnusson, D. (1991). *Theoretical Foundations of Psychological Tests* (M. N. Baraheni, Trans.). Tehran: University of Tehran. [Persian]

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge.

Mortaz Hejri, S., Khabaz Mafinezhad, M., & Jalili, M. (2014). Guessing in multiple choice questions: Challenges and strategies. *Iranian Journal of Medical Education, 14*(7), 594–604. [Persian]

Narayanon, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement, 20*(3), 257–274.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal, 5*(2), 107–124.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*(3), 517–528.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114*(3), 552–566.

Ritz, C., & Streibig, J. C. (2008). *Nonlinear regression with R* (Vol. 10). Springer.

Saif, A. A. (2018). *Measurement, Assessment, and Educational Evaluation*. Tehran: Doran Publishing. [Persian]

Santelices, M. V., & Wilson, M. (2012). On the relationship between differential item functioning and item difficulty: An issue of methods? Item response theory approach to differential item functioning. *Educational and Psychological Measurement, 72*(1), 5–36.

Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice, 23*(2), 226–235.

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). Lawrence Erlbaum Associates.

Tjalla, A., & Fitriani, S. (2017). The effect of multiple choice scoring methods and risk taking attitude toward chemistry learning outcomes. *Jurnal Pendidikan Dan Kebudayaan*, 2(2), 199–210.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*(2), 139–158.

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement, 73*(3), 532–547.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa: National Defense Headquarters.