

Evaluating of Gender Differences in English Language Items on the Master of Psychology Series Exam Using the Wald Test, Likelihood Ratio Test, and Revised Likelihood Ratio Test Based on the Cognitive Diagnostic Model

Shoeayb Qasemi  *

Corresponding Author, Ph.D. Student of Educational Measurement and Assessment, Allameh Tabataba'i University, Tehran, Iran. E-mail: q.shoeayb@gmail.com

Ali Delavar 

Professor, Department of Educational Measurement and Assessment, Allameh Tabataba'i University, Tehran, Iran. E-mail: delavarali@yahoo.com

Farzad Eskandari 

Professor Department of Statistics, Allameh Tabataba'i University, Tehran, Iran. E-mail: ffeskandari@yahoo.com

Abstract

This research was carried out to investigate the presence of gender-related differences in English language items found in the Master of Psychology Series exam, employing cognitive diagnostic model-based methodologies. This study utilized a sample population of 2,455 female applicants and 919 male applicants who participated in the Master of Psychology series exam held in 1396. The G-DINA cognitive diagnostic model was applied to fit the data. Further analysis was conducted to detect differential item function (DIF) in the questions, employing the Wald test, likelihood ratio test, and revised likelihood ratio test methodologies. The findings from all three methods demonstrated moderate agreement regarding the identification of differential item functioning (DIF). A total of 16 questions were identified as having DIF. Notably, most of the questions indicated consistent DIF favoring males. Questions found to have non-uniform DIF demonstrated that those in favor of men were more pronounced among the dominant applicants, while the opposite holds true for non-dominant applicants. To further assess the presence of bias and the causes behind DIF, it is suggested that additional studies be carried out in this area.

Keywords: Differential Item Functioning (DIF); Gender; Cognitive Diagnostic Model

Cite this Article: Qasemi, S., Delavar, A., & Eskandari, F. (2025). Evaluating of Gender Differences in English Language Items on the Master of Psychology Series Exam Using the Wald Test, Likelihood Ratio Test, and Revised Likelihood Ratio Test Based on the Cognitive Diagnostic Model. *Educational Measurement*, 16(59), 37-61. <https://doi.org/10.22054/jem.2024.73600.3460>



© 2016 by Allameh Tabataba'i University Press
Publisher: Allameh Tabataba'i University Press

Introduction

Over the last few decades, the evaluation of differential item functioning (DIF) has become an integral part of the test item analysis process. A significant number of cognitive tests possess some degree of multidimensionality. The deficiency of DIF detection methods in capturing the complete spectrum of constructs can result in bias in the assessment of DIF in cognitive tests. Cognitive diagnostic models, as multidimensional models, provide a more comprehensive coverage of the underlying space relative to unidimensional models, as noted by Zhang (2006).

Despite the growing popularity of cognitive diagnostic models, multiple aspects surrounding their application remain to be explored further. One such aspect concerns the study of differential item functioning (DIF). The potential presence of DIF in cognitive diagnostic models can yield biased and inaccurate parameter estimates and examinee proficiency profiles. Skill profiles estimated in the presence of items with DIF can lead to flawed decisions concerning the mastery of particular skills by individuals within the focal or reference group (Hou et al., 2014). While the majority of research interest has primarily focused on the Wald test, there has been limited exploration of the likelihood ratio test within the context of cognitive diagnostic models. Consequently, a more extensive investigation into the performance of this method, compared to the widely used Wald test, appears warranted.

Research Question(s)

The objectives of this study are twofold: 1) To determine if the findings from the likelihood ratio test and the revised likelihood ratio test align with those of the Wald test in identifying differential item functioning (DIF), and 2) To evaluate the status of English language items on the Master of Psychology series exam in terms of DIF based on gender.

Literature Review

Various methods have been put forth for identifying DIF in cognitive diagnostic models. One such study by Zhang (2006) scrutinized the effectiveness of the Mixture-Homogeneity (MH) and SIBTEST methods for detecting DIF using the DINA model, employing matched examinee pairs based on their proficiency profile scores. Li and Baron

(2008) built upon the DINA model to propose a higher-order modified version capable of concurrently identifying both differential item and attribute functioning. Furthermore, Li and Wang (2015) introduced a general DIF method known as LCDM-DIF, designed for utilization with a wide array of cognitive diagnostic models, involving more than two groups of examinees and diverse grouping variables, including both categorical and continuous factors.

A notable limitation of higher-order DINA and LCDM-DIF methods is the considerable time expenditure required for parameter estimation, which can take several hours. The Wald test, specifically CDM-WALD, has been suggested for assessing DIF in cognitive diagnostic models and has demonstrated satisfactory performance in various studies (George & Robitzsch, 2014; Hou et al., 2014). Additionally, the log-likelihood ratio test (CDM-LR) and its modified counterpart have been proposed to assess DIF in cognitive diagnostic models (Ma et al. 2017; 2021).

Methodology

The research population encompassed all candidates taking the Master of Psychology series exam in the year 2017. The sample included 3,374 candidates, with 2,455 females (73%) and 919 males (27%). The measurement tool utilized for this study was the English language Master of Psychology series exam with code 1133 held in 2017, which consisted of 30 items divided into 5 passages of reading comprehension, each containing six items.

The data analysis commenced with the identification of cognitive skills and the formation of a Q-matrix. Following this, the G-DINA cognitive diagnostic model was chosen for the current study utilizing the log-likelihood ratio test. In the ultimate phase, DIF was evaluated for the items on the exam. The analyses were carried out employing the GDINA package (Ma et al., 2018) and psych package (Revelle, 2018) in the R software.

Results

The Q-matrix was established based on the skills of vocabulary, grammar, explicit information extraction, integration and synthesis, and reasoning. This finding aligned with previous studies (Jang, 2009; Ravand et al., 2012; Li, 2011) that similarly utilized these five skills. A detailed examination of item parameters revealed high guess and

slipping parameters, indicative of the generally poor item quality. The Wald test identified a total of 13 items with differential item functioning (DIF). The revised likelihood ratio test and the likelihood ratio test identified 11 and 5 items, respectively, as having DIF. The Kappa coefficient between the Wald test and the likelihood ratio test was 0.46, and between the likelihood ratio test and the revised likelihood ratio test, it was 0.51.

Conclusion

The Likelihood ratio test emerged as the most conservative method in identifying DIF, with all five items identified by this test also being identified by the other two methods. Previous studies (Ma et al., 2017; Ma et al., 2021) have demonstrated that the revised Likelihood ratio test performed effectively under diverse simulation scenarios, especially in instances with poor item quality and moderate to large sample sizes. Notably, it demonstrated the best performance in terms of Type I error. The high guess and slipping parameters of the English language exam items could contribute to the relatively large number of DIF items identified by the Wald test.

The Kappa coefficient revealed a moderate level of agreement among the three methods in identifying DIF. Among these methods, the highest agreement was found between the revised likelihood ratio test and the likelihood ratio test. The Likelihood ratio test exhibited effectiveness in identifying DIF within cognitive diagnostic models, with results that were comparable to those derived from the Wald test.

The analysis revealed that a total of 16 items exhibited differential item functioning (DIF) based on gender. Out of these items, 11 were identified as having uniform DIF, with an advantage for male candidates. The occurrence of DIF poses a significant threat to exam validity, as it introduces errors and potential unfairness in decisions related to student admissions, with female examinees being disproportionately affected in the Master of Psychology entrance exam.

بررسی کنش افتراقی سؤالات زبان انگلیسی آزمون کارشناسی ارشد روان‌شناسی نسبت به جنسیت به‌منظور مقایسه آزمون والد، آزمون نسبت درستنمایی و نسبت درستنمایی بازبینی شده در مدل تشخیصی شناختی

شعب قاسمی *

نویسنده مسئول، دانشجوی دکتری رشته سنجش و اندازه‌گیری، دانشگاه علامه طباطبائی،
تهران، ایران. رایانامه: q.shoeayb@gmail.com

علی دلاور

استاد گروه سنجش و اندازه‌گیری، دانشگاه علامه طباطبائی، تهران، ایران. رایانامه:
delavarali@yahoo.com

فرزاد اسکندری

استاد گروه آمار، دانشگاه علامه طباطبائی، تهران، ایران. رایانامه: askandari@atu.ac.ir

چکیده

این پژوهش با هدف مقایسه روش‌های شناسایی کنش افتراقی سؤال در مدل‌های تشخیصی شناختی بر روی سؤالات درس زبان انگلیسی آزمون کارشناسی ارشد مجموعه روان‌شناسی بر اساس جنسیت انجام شد. نمونه تحقیق شامل ۲۴۵۵ داوطلب زن و ۹۱۹ داوطلب مرد بود که در آزمون کارشناسی ارشد مجموعه روان‌شناسی سال ۱۳۹۶ شرکت کرده بودند. مدل تشخیصی شناختی جی‌دینا با داده‌ها برازش داده شد. کنش افتراقی سؤالات با سه روش آزمون والد، آزمون نسبت درستنمایی و آزمون نسبت درستنمایی بازبینی شده بررسی شد. نتایج این سه روش توافق متوسطی در تشخیص کنش افتراقی سؤال داشتند. دو روش مبتنی بر درستنمایی می‌توانند در کنار روش والد برای شناسایی کنش افتراقی سؤال به کار روند. بر اساس نتایج ۱۶ سؤال دارای کنش افتراقی بودند. بیشتر سؤالات دارای کنش افتراقی سؤال یکنواخت به نفع مردان بودند. سؤالات دارای کنش افتراقی غیریکنواخت برای افراد مسلط به نفع مردان و برای افراد غیرمسلط به نفع زنان عمل می‌کردند. توصیه می‌شود مطالعات تکمیلی برای بررسی وجود سوگیری و علت آن در سؤالات دارای کنش افتراقی سؤال انجام شود.

کلیدواژه‌ها: کنش افتراقی سؤال، جنسیت، زبان انگلیسی، مدل تشخیصی شناختی

استناد به این مقاله: شعب قاسمی، دلاور، علی، و اسکندری، فرزاد. (۱۴۰۴). بررسی کنش افتراقی سؤالات زبان انگلیسی آزمون کارشناسی ارشد روان‌شناسی نسبت به جنسیت به‌منظور مقایسه آزمون والد، آزمون نسبت درستنمایی و نسبت درستنمایی بازبینی شده در مدل تشخیصی شناختی مقاله. فصلنامه اندازه‌گیری تربیتی، ۱۶(۵۹)، ۳۷-۶۱.
<https://doi.org/10.22054/jem.2024.73600.3460>

مقدمه

ارزیابی کنش افتراقی سؤال^۱ (DIF) در دهه‌های اخیر به بخشی از روال تحلیل سؤال تبدیل شده است. کنش افتراقی سؤال در ادبیات روان‌سنجی توجه زیادی دریافت کرده است زیرا با عادلانه بودن و روایی آزمون ارتباط مستقیم دارد. یک سؤال زمانی دارای کنش افتراقی سؤال است که احتمال پاسخ صحیح به سؤال آزمودنی‌های گروه مرجع و کانونی که توانایی یکسانی دارند متفاوت باشد. تشخیص دادن کنش افتراقی سؤال در سؤال‌های یک آزمون از نظر کیفیت ابزار ارزیابی دارای اهمیت است، زیرا کنش افتراقی سؤال می‌تواند به عنوان وجود ابعاد مزاحم^۲ توصیف شود که با توانایی یا خصیصه‌ای که قصد اندازه‌گیری آن را داریم تداخل دارند (Ackerman, 1992).

تکنیک‌های استاندارد ارزیابی کنش افتراقی سؤال به ساختار تک‌بعدی آزمون متکی هستند. زمانی که مفروضه تک‌بعدی بودن برقرار است، هم‌تاسازی آزمودنی‌ها بر اساس یک نمره کل منطقی است. اما وقتی ساختار آزمون تک‌بعدی نیست، هم‌تاسازی سنتی با استفاده از نمره کل آزمون و یک توانایی مکنون احتمال ارتکاب خطای نوع اول را افزایش می‌دهد. بسیاری از آزمون‌های شناختی تا حدودی چندبعدی هستند. ناتوانی روش‌های تشخیص کنش افتراقی سؤال در به حساب آوردن فضای کامل مکنون می‌تواند منجر به سوگیری در ارزیابی کنش افتراقی سؤال در آزمون‌های شناختی شود (Zhang, 2006).

مدل‌های تشخیصی شناختی^۳ مدل‌های چندبعدی هستند، این مدل‌ها در مقایسه با مدل‌های تک‌بعدی، فضای مکنون را بهتر پوشش می‌دهند (Zhang, 2006). مدل‌های تشخیصی شناختی پتانسیل بالایی برای یادگیری و تدریس دارند. در مدل‌های تشخیصی شناختی برخلاف نظریه کلاسیک و نظریه سؤال-پاسخ به جای یک نمره، برای هر پاسخ‌دهنده یک پروفایل برآورد می‌شود که بیانگر تسلط یا عدم تسلط او بر مهارت‌هایی است که آزمون اندازه می‌گیرد. برای مثال در یک آزمون که سه مهارت را اندازه می‌گیرد پروفایل $\alpha=(0.110)$ نشان می‌دهد که آزمودنی به مهارت‌های دوم و سوم مسلط شده است اما به مهارت‌های اول و چهارم مسلط نشده است. مهارت‌های مورد نیاز برای پاسخگویی به هر سؤال توسط ماتریس Q مشخص می‌شود. ردیف‌های ماتریس Q نشان‌دهنده سؤال‌ها و ستون

1. differential item functioning

2. nuisance

3. cognitive diagnostic model

های آن نشان‌دهنده مهارت‌های مورداندازه‌گیری هستند. چنانچه پاسخ صحیح به یک سؤال نیازمند یک مهارت شناختی باشد، سلول مرتبط با آن در ماتریس عدد «۱» و در غیر این صورت عدد «۰» خواهد بود. جدول ۱ یک نمونه ماتریس Q را نشان می‌دهد. پاسخ صحیح به سؤال یک نیازمند تسلط به مهارت ۱ است و مهارت‌های ۲ و ۳ نقشی در پاسخ صحیح به این سؤال ندارند. برای سؤال ۲ پاسخ صحیح نیازمند تسلط به مهارت‌های ۲ و ۳ است (Hunter et al., 2022).

جدول ۱. نمونه ماتریس Q

مهارت ۱	مهارت ۲	مهارت ۳
سؤال ۱	۰	۰
سؤال ۲	۰	۱

مدل‌های تشخیصی شناختی (CDM) در سال‌های اخیر موردعلاقه محققان قرار گرفته‌اند اما با وجود توجه روزافزون به مدل‌های تشخیصی شناختی می‌توان گفت این حیطه هنوز در مرحله نوزادی قرار دارد (Leighton & Gierl, 2007). هنوز مسائل بسیاری در مورد مدل‌های تشخیصی شناختی باید موردبررسی قرار گیرد. یکی از این مسائل بررسی کنش افتراقی سؤال است (Hou, 2013؛ Hou et al., 2014). وجود کنش افتراقی سؤال در مدل‌های تشخیصی شناختی ممکن است به برآوردهای پارامتر سؤال و برآوردهای پروفایل خصیصه مکنون نامعتبر بیانجامد. پروفایل‌های مهارت برآورد شده در صورت وجود سؤال‌های دارای کنش افتراقی سؤال می‌تواند موجب تصمیم‌گیری اشتباه در مورد مهارت‌هایی شود که اعضای گروه کانونی یا مرجع به آن‌ها تسلط یافته‌اند (Hou et al., 2014). پژوهش Paulsen و همکاران (2020) نشان داده صحت طبقه‌بندی پروفایل‌ها به گونه منفی تحت تأثیر کنش افتراقی قرار می‌گیرد.

در بافت مدل‌های تشخیصی شناختی زمانی یک سؤال دارای کنش افتراقی است که احتمال پاسخ صحیح به آن سؤال برای آزمودنی‌های با پروفایل یکسان (یعنی افرادی که تسلط یا عدم تسلط آن‌ها بر روی مهارت‌های مورداندازه‌گیری یکسان است) اما متعلق به گروه‌های مختلف یکسان نباشد. چنانچه احتمال پاسخ صحیح به سؤال برای آزمودنی‌های عضو گروه کانونی (F) و مرجع (R) که دارای پروفایل یکسانی هستند برابر باشد سؤال فاقد کنش افتراقی سؤال است. کنش افتراقی سؤال در مدل‌های تشخیصی شناختی را می‌توان به

این صورت نشان داد: $\Delta_j \alpha_l = P(X_j = 1 | \alpha_l)_F - P(X_j = 1 | \alpha_l)_R$ ؛ $\Delta_j \alpha_l$ نشان‌دهنده کنش افتراقی سؤال در سؤال j برای آزمودنی‌های دارای پروفایل تسلط α_l است. $P(X_j = 1 | \alpha_l)_F$ احتمال پاسخ صحیح به سؤال j برای آزمودنی‌های عضو گروه F است که دارای پروفایل α_l هستند. $P(X_j = 1 | \alpha_l)_R$ احتمال پاسخ صحیح به سؤال j برای آزمودنی‌های عضو گروه R است که دارای پروفایل α_l هستند. زمانی که $\Delta_j \alpha_l > 0$ باشد، کنش افتراقی سؤال به نفع گروه کانونی و به ضرر گروه مرجع است. زمانی که $\Delta_j \alpha_l < 0$ باشد، کنش افتراقی سؤال به نفع گروه مرجع و به ضرر گروه کانونی است. زمانی که برای همه پروفایل‌ها $\Delta_j \alpha_l = 0$ است سؤال فاقد کنش افتراقی است. چنانچه $\Delta_j \alpha_l$ برای همه پروفایل‌های ممکن مثبت یا منفی باشد کنش افتراقی سؤال یکنواخت است. چنانچه علامت $\Delta_j \alpha_l$ پروفایل‌های مختلف تغییر کند، کنش افتراقی سؤال غیریکنواخت است (Hou, 2013; Hou et al., 2014).

چند روش برای شناسایی کنش افتراقی در مدل‌های تشخیصی شناختی پیشنهاد شده است. تفاوت‌های گروهی در پروفایل تسلط مهارت بعد از کنترل توانایی کلی در چهارچوب مدل تشخیصی شناختی بررسی شده است (Milewski & Baron, 2002). هم‌تاسازی آزمودنی‌ها بر اساس نمره پروفایل خصیصه آن‌ها با استفاده از مدل DINA بررسی شده تا کارایی روش‌های منتل-هنسزل^۱ (MH) و SIBTEST برای تشخیص کنش افتراقی سؤال را بررسی شود (Zhang, 2006). این دو روش خطای نوع اول بالا و توان آماری پایینی داشتند. مطالعه میلوسکی و بارون بسط داده شده (Li, 2008) و از مدل DINA^۲ مرتبه بالاتر اصلاح شده برای تشخیص هم‌زمان کنش افتراقی سؤال و کنش افتراقی خصیصه استفاده شده است. یک روش ارزیابی کنش افتراقی سؤال کلی بر اساس مدل LCDM پیشنهاد شده (LCDM-DIF) که برای مدل‌های تشخیصی شناختی مختلف، بیشتر از دو گروه از آزمودنی‌ها و متغیرهای گروه‌بندی طبقه‌ای، پیوسته، مشاهده شده و ممکن کاربرد دارد (Li & Wang, 2015). یکی از مشکلات روش‌های مدل مرتبه بالاتر DINA و LCDM-DIF زمان‌بر بودن فرایند برآورد پارامترها است که تا چند ساعت زمان نیاز دارد. آزمون والد برای ارزیابی کنش افتراقی سؤال در مدل‌های تشخیصی شناختی (CDM-WALD) پیشنهاد شده است (Hou, 2013; Hou et al., 2014). این آزمون در چند مطالعه برای ارزیابی کنش افتراقی

1. Mantel-Hanszel

2. deterministic inputs, noisy and gate

سؤال استفاده شده است و نشان داده است که عملکرد قابل قبولی دارد (George & Robitzsch, 2014؛ کاظمی دانا، ۱۳۹۵؛ Svetina et al., 2018؛ سعادت و همکاران، ۱۳۹۹). آزمون والد برای شناسایی کنش افتراقی سؤال در سنجش انطباقی کامپیوتری تشخیص شناختی نیز مورد استفاده قرار گرفته است و خطای نوع اول آن نزدیک به مقدار اسمی بوده است (Wang & Wang, 2022). آزمون نسبت درستنمایی^۱ (CDM-LR) و نسخه اصلاح شده آن نیز برای ارزیابی کنش افتراقی سؤال در مدل‌های تشخیصی شناختی پیشنهاد شده‌اند. کاربرد روش نسبت درستنمایی در مدل‌های تشخیصی شناختی برخلاف آماره والد نیازمند برآورد دقیق پارامترهای سؤال و ماتریس واریانس-کوواریانس مجانب سؤال نیست. روش‌های برآورد سریع برای مدل‌های تشخیصی شناختی چندگروهی که برای آزمون نسبت درستنمایی ضروری هستند در دسترس قرار دارند، و در چند ثانیه به همگرایی می‌رسند (Ma et al., 2017؛ Ma et al., 2021). آزمون والد و آزمون نسبت درستنمایی برای شناسایی کنش افتراقی در مدل‌های تشخیصی شناختی چندارزشی نیز مورد استفاده قرار گرفته‌اند و در مقایسه با روش‌هایی که مبتنی بر مدل نیستند عملکرد بهتری نشان داده‌اند (Qui-Yun., 2022).

بررسی پیشینه پژوهشی نشان می‌دهد بیشتر توجه پژوهشی معطوف به آزمون والد بوده است و تحقیقات اندکی بر روی آزمون نسبت درستنمایی برای شناسایی کنش افتراقی سؤال در مدل‌های تشخیصی شناختی انجام شده است. لذا انجام تحقیقات بیشتر برای مقایسه عملکرد این روش با آزمون والد که پرکاربردترین روش در این حوزه است ضروری به نظر می‌رسد. در این تحقیق عملکرد این روش‌ها روی داده‌های واقعی مقایسه می‌شود. درس زبان انگلیسی آزمون کارشناسی ارشد مجموعه روان‌شناسی برای این منظور انتخاب شد و کنش افتراقی سؤال نسبت به جنسیت بررسی شد. انتخاب زبان انگلیسی آزمون کارشناسی ارشد به دو دلیل بود: نخست بحث عادلانه بودن آزمون‌ها در آزمون‌های سرنوشت‌ساز مانند کنکور کارشناسی ارشد اهمیتی دوچندان پیدا می‌کند و نتایج حاصل از این مطالعه بازخوردی در اختیار طراحان سؤالات درس زبان انگلیسی کنکور کارشناسی ارشد مجموعه روان‌شناسی می‌گذارد که یاریگر آنان در بهبود عادلانه بودن آزمون خواهد شد. دوم این که بازبرازش^۲ مدل‌های تشخیصی شناختی با آزمون‌های موجود امر رایجی است تا هم ساختار مهارت‌های

1. likelihood ratio test

2. retrofitting

مکنون آن‌ها شناسایی شود و هم اطلاعات تشخیصی در مورد عملکرد پاسخ‌دهنده‌ها به دست آید. به علت دشواری و زمان‌بر بودن فرایند ساخت آزمون، تاکنون آزمون‌های تشخیصی واقعی اندکی ساخته شده‌اند (Alderson & Huhta, 2005؛ Urmston et al, 2013). بیشتر کاربردهای مدل‌های تشخیصی شناختی به بازبرازش این مدل‌ها با آزمون‌های موجود اختصاص داشته است. چند مطالعه به‌طور موفقیت‌آمیز مدل‌های تشخیصی شناختی را برای آزمون‌های درک مطلب زبان انگلیسی به کار برده‌اند (برای مثال، Wand & Jang, 2009؛ Gierl, 2011؛ Ravand et al., 2012؛ Yi, 2012؛ Li et al., 2015؛ Effatpanah et al., 2019)؛ بنابراین این تحقیق در پی پاسخ دادن به دو سؤال است:

نتایج آزمون نسبت درست‌نمایی و نسبت درست‌نمایی بازبینی‌شده در شناسایی کنش افتراقی سؤال با آزمون والد همخوان است؟

وضعیت سؤالات درس زبان انگلیسی کارشناسی ارشد رشته روان‌شناسی از نظر کنش افتراقی سؤال نسبت به جنسیت چگونه است؟

در ادامه ابتدا آزمون والد، آزمون نسبت درست‌نمایی و آزمون نسبت درست‌نمایی اصلاح شده شرح داده شده است و سپس به روش‌شناسی و یافته‌های تحقیق پرداخته شده است. آزمون والد: فرایند آزمون فرض Wald-CDM را می‌توان به‌صورت ماتریس تشکیل داد. در مدل G-DINA ۱ اولین مرحله از روش Wald-CDM این است که پارامترهای سؤال به‌طور جداگانه برای گروه‌های F و R برآورد شوند که در قالب بردار به‌صورت زیر نوشته می‌شوند:

$$\hat{\beta}_j^* = (\hat{\beta}_{Rj}, \hat{\beta}_{Fj})' = (\hat{P}(a_{0j}^*)_R, \dots, \hat{P}(a_{lj}^*)_R, \dots, \hat{P}(a_{1j}^*)_R, \hat{P}(a_{0j}^*)_F, \dots, \hat{P}(a_{lj}^*)_F, \dots, \hat{P}(a_{1j}^*)_F)',$$

عناصر بردار نشان‌دهنده پارامترهای سؤال مدل برای گروه‌های F و R است. در مرحله دوم، فرض صفر برابری پارامترهای سؤال، یعنی $\beta_j = 0$ ، $H_0: R_j$ ، برای گروه‌های F و R آزمایش می‌شود. از آنجا که تعداد پارامترهایی که برای هر گروه باید برآورد شوند 2^{K_j} است، 2^{K_j} محدودیت وجود دارد و ابعاد ماتریس محدودیت R_j برابر $2^{K_j} \times 2^{K_j+1}$ است. به‌عنوان مثال، برای یک سؤال که برای پاسخ صحیح نیاز به دو صفت دارد ($K_j = 2$)، ماتریس R_j به‌صورت زیر است:

$$\mathbf{R}_j = \begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix}.$$

فرضیه صفر می‌تواند در قالب مدل محدود به شرح زیر بیان شود:

$$H_0 : \mathbf{R}_j \cdot \boldsymbol{\beta}_j^* = 0$$

آماره Wald یعنی W_j برای آزمون H_0 به این صورت محاسبه می‌شود:

$$W_j = [\mathbf{R}_j \cdot \boldsymbol{\beta}_j^*]' \{ \mathbf{R}_j \cdot \text{Var}(\boldsymbol{\beta}_j^*) \cdot \mathbf{R}_j' \}^{-1} [\mathbf{R}_j \cdot \boldsymbol{\beta}_j^*].$$

تحت فرض صفر که $\beta_j=0$ است، فرض می‌شود که W_j توزیع مجانبی با درجات آزادی 2^k دارد. اگر W_j از مقدار معناداری توزیع χ^2 برای یک سطح معناداری معین بیشتر باشد، فرض صفر رد می‌شود. لازم به ذکر است که روش Wald-CDM فقط به برآورد مدل بدون محدودیت نیاز دارد. با مدرج سازی‌های جداگانه برای گروه‌های F و R ، مشکل آلودگی برآوردهای صفت و پارامترهای سؤال به خاطر سؤال‌های بالقوه دارای کنش افتراقی سؤال، از بین می‌رود (Hou, 2013؛ Hou et al., 2014؛ Hou et al., 2020).

آزمون CDM-LR: در آزمون CDM-LR (Ma et al., 2017؛ Ma et al., 2021) ابتدا یک مدل چندگروهی با محدودیت یکسان بودن پارامترهای همه سؤال‌ها در همه گروه‌ها با داده‌ها برازش داده می‌شود. مقدار لگاریتم درستنمایی این مدل را با $l_{allequal}$ نشان داده می‌شود. در مرحله بعد یک مدل تشخیصی شناختی چندگروهی که به جز سؤال مورد مطالعه سایر سؤال‌ها دارای محدودیت یکسان بودن پارامترهای سؤال در همه گروه‌ها هستند با داده‌ها برازش داده می‌شود. مقدار لگاریتم درستنمایی این مدل را با $l_{jnotequal}$ نشان می‌دهیم. آماره $G_j^2 = -2[l_{allequal} - l_{jnotequal}]$ محاسبه می‌شود. این آماره دارای توزیع کای اسکوتر با درجه آزادی m است که m تفاوت تعداد پارامترهای دو مدل است. در مدل‌های تشخیصی شناختی امکان محاسبه آزمون نسبت درستنمایی به یک شیوه دیگر

وجود دارد که آزمون نسبت درستنمایی بازبینی شده^۱ (R-CDM-LR) نام دارد. در این روش ابتدا یک مدل تشخیصی شناختی چند گروهی بدون محدودیت یکسان بودن پارامترهای سؤال در گروه‌های مختلف با داده‌ها برازش داده می‌شود. لگاریتم درستنمایی این مدل با $L_{lunequal}$ نشان داده می‌شود. سپس یک مدل تشخیصی شناختی چند گروهی با داده‌ها برازش داده می‌شود که سؤال مورد مطالعه دارای محدودیت یکسان بودن پارامترها در گروه‌های مختلف است. لگاریتم درستنمایی این مدل به صورت L_{jequal} نشان داده می‌شود. آماره $G^2_{df} = -2[l_{lunequal} - l_{jequal}]$ محاسبه می‌شود.

روش

تحقیق حاضر یک پژوهش توصیفی است که از نظر هدف در زمره تحقیقات کاربردی قرار می‌گیرد. از این نظر که تحقیق حاضر بر روی داده‌های درس زبان انگلیسی آزمون کارشناسی ارشد مجموعه روان‌شناسی انجام می‌شود، می‌توان تحلیل از نوع تحلیل‌های ثانویه در نظر گرفت.

جامعه آماری و نمونه: جامعه پژوهش شامل پاسخ‌های کلیه داوطلبان آزمون کارشناسی ارشد مجموعه روان‌شناسی در سال ۱۳۹۶ بود. نمونه تحقیق ۳۳۷۴ داوطلب را در بر می‌گرفت که متشکل از ۲۴۵۵ داوطلب زن (۷۳ درصد) و ۹۱۹ داوطلب مرد (۲۷ درصد) بود. با توجه به حجم بالای نمونه در هر دو گروه، امکان برآورد پارامترهای مدل در هر گروه وجود دارد و تفاوت حجم نمونه مشکلی ایجاد نمی‌کند.

ابزار اندازه‌گیری: ابزار اندازه‌گیری مورد استفاده در این پژوهش آزمون زبان انگلیسی کارشناسی ارشد مجموعه روان‌شناسی کد ۱۱۳۳ در سال ۱۳۹۶ بود. این آزمون ۳۰ سؤالی متشکل از ۵ متن درک مطلب است که هر متن دارای ۶ سؤال است. پاسخ صحیح به هر سؤال نمره ۱ و پاسخ نادرست نمره صفر دریافت کرد. برای سؤال‌های بدون پاسخ که داوطلب به آن‌ها پاسخ نداده و سؤال‌های جامانده که داوطلب زمان لازم برای پاسخگویی به آن‌ها را نیافته نمره صفر در نظر گرفته شد.

تحلیل داده‌ها: نخست مهارت‌های شناختی آزمون تعیین و Q تشکیل شد. پس از تدوین ماتریس Q با استفاده از آزمون نسبت درستنمایی مدل تشخیصی شناختی G-DINA برای تحقیق حاضر انتخاب شد. در مرحله آخر کنش افتراقی سؤال سؤالات آزمون شد. تحلیل‌ها

1. revised CDM-LRDIF

با استفاده از بسته‌های GDINA (Ma et al., 2018) و psych (Revelle, 2018) در نرم‌افزار R انجام شد.

یافته‌ها

با توجه به اینکه مطالعه حاضر بازبرازش^۱ یک مطالعه تشخیصی شناختی با آزمونی که با این هدف طراحی نشده است، مرحله اول تعیین مهارت‌های شناختی مورداندازه‌گیری توسط آزمون است. بررسی ادبیات تحقیق در زمینه کاربرد مدل‌های تشخیصی شناختی برای آزمون‌های درک مطلب زبان انگلیسی نشان داد که ۵ مهارت شناختی زیربنایی در این آزمون‌ها قابل شناسایی هستند: واژگان^۲، دستور زبان^۳، استخراج اطلاعات صریح^۴، اتصال و ترکیب کردن^۵، و استنتاج^۶ (Jang, 2009؛ Ravand et al., 2012؛ Li, 2011). در این مطالعه نیز ۵ مهارت یادشده به‌عنوان مهارت‌های زیربنایی آزمون زبان کارشناسی ارشد مجموعه روان‌شناسی انتخاب شدند.

برای تدوین ماتریس Q، سؤالات آزمون زبان انگلیسی کارشناسی ارشد روان‌شناسی سال ۱۳۹۶ به همراه لیست مهارت‌های شناختی ۵ گانه همراه به دو متخصص زبان انگلیسی داده شد تا مهارت‌های لازم برای پاسخگویی به هر سؤال را مشخص کنند، به هر دو متخصص توضیحات لازم در زمینه مهارت‌های شناختی پنج‌گانه داده شد. دو متخصص به‌صورت مستقل از یکدیگر به تدوین ماتریس Q پرداختند. ماتریس‌های Q تدوین‌شده توسط دو متخصص در مرحله بعد توسط پژوهشگران و یک متخصص آموزش زبان انگلیسی موردبررسی قرار گرفت و پس از انطباق و اصلاح آن‌ها ماتریس Q پیشنهادی تهیه شد. روایی این ماتریس پیشنهادی با روش دلاتوره و چو (De la Torre & Chiu, 2016) بررسی شد. بر اساس نتایج به‌دست‌آمده درایه‌های ماتریس Q برای سؤال‌های شماره ۵، ۱۱، ۱۴، ۲۷، و ۲۹ مورد اصلاح قرار گرفت. ماتریس Q اصلاح‌شده به‌عنوان ماتریس Q نهایی در تحلیل داده‌ها استفاده شد (جدول ۱).

-
1. retrofit
 2. vocabulary
 3. syntax
 4. extracting explicit information
 5. connecting and synthesizing
 6. making inferences

جدول ۱. ماتریس Q سؤال‌های زبان انگلیسی

سؤال	واژگان	دستور زبان	استخراج اطلاعات صریح	اتصال و ترکیب	استنتاج
۱	۰	۱	۱	۰	۰
۲	۱	۰	۰	۰	۰
۳	۱	۰	۰	۰	۱
۴	۱	۱	۰	۱	۰
۵	۱	۰	۰	۱	۰
۶	۰	۱	۰	۰	۱
۷	۰	۰	۰	۱	۰
۸	۱	۱	۰	۱	۰
۹	۱	۰	۰	۰	۰
۱۰	۱	۰	۰	۱	۰
۱۱	۰	۱	۱	۰	۰
۱۲	۱	۰	۰	۱	۰
۱۳	۱	۰	۰	۱	۰
۱۴	۱	۰	۰	۱	۰
۱۵	۱	۰	۰	۰	۰
۱۶	۰	۱	۰	۱	۰
۱۷	۱	۰	۰	۱	۰
۱۸	۱	۰	۰	۰	۱
۱۹	۰	۱	۰	۰	۱
۲۰	۱	۰	۰	۱	۰
۲۱	۱	۱	۰	۱	۰
۲۲	۰	۰	۱	۰	۰
۲۳	۱	۰	۰	۰	۰
۲۴	۰	۱	۰	۱	۰
۲۵	۰	۱	۰	۱	۰
۲۶	۰	۱	۱	۰	۰
۲۷	۰	۰	۱	۰	۰
۲۸	۱	۰	۰	۰	۰
۲۹	۱	۰	۰	۱	۰
۳۰	۱	۰	۰	۱	۰

جدول ۲. شاخص‌های برازش مدل‌های تشخیصی شناختی

مدل	تعداد پارامترها	-2LL	AIC	BIC	کای اسکوتر	درجه آزادی	سطح معناداری
GDINA	۱۵۱	-۵۳۷۳۳/۰۹	۱۰۷۷۶۸/۱۷	۱۰۸۶۹۲/۸۷	-	-	-
DINA	۹۱	-۵۴۶۶۴/۹۵	۱۰۹۵۱۱/۹۰	۱۱۰۰۶۹/۱۷	۱۸۶۳/۷۳	۶۰	۰/۰۰۱
RRUM	۱۱۷	-۵۴۰۹۲/۱۷	۱۰۸۴۱۸/۳۳	۱۰۰۱۳۴/۸۲	۷۱۸/۱۶	۳۴	۰/۰۰۱
DINO	۹۱	-۵۴۶۷۵/۸۹	۱۰۹۵۳۳/۷۸	۱۱۰۰۹۱/۰۵	۱۸۸۵/۶۰	۶۰	۰/۰۰۱
ACDM	۱۱۷	-۵۴۱۰۰/۲۸	۱۰۸۴۳۴/۵۶	۱۰۹۱۵۱/۰۵	۷۳۴/۳۹	۳۴	۰/۰۰۱

برای انتخاب مدل شاخص‌های برازش مدل‌های مختلف مورد مقایسه قرار گرفتند (به جدول ۲ نگاه کنید). مقدار $-2LL$ ^۱، AIC ^۲، و BIC ^۳ به دست آمده برای مدل GDINA کمتر از سایر مدل‌ها است. آزمون نسبت درستمایی نشان داد که مدل‌های DINA^۴، DINO^۴، RRUM^۵ و ACDM^۶ در مقایسه با مدل GDINA برازش ضعیف‌تری با داده‌ها دارند ($P < ۰/۰۰۱$)؛ بنابراین مدل GDINA انتخاب و با استفاده از ماتریس Q به دست آمده با داده‌ها برازش داده شد.

پارامترهای سؤالات برای گروه زنان و مردان در قالب دو پارامتر حدس^۷ (g) و لغزش^۸ (s) در جدول ۳ آورده شده است. در مدل G-DINA پارامتر حدس همان عرض از مبدأ δ_{j0} است، و پارامتر لغزش برابر با $s_j = 1 - P(\alpha_{ij}^*)$ است. پارامتر حدس نشان‌دهنده احتمال پاسخ صحیح به سؤال برای آزمودنی است که فاقد مهارت‌های لازم برای پاسخ صحیح به سؤال است. پارامتر لغزش نشان‌دهنده احتمال پاسخ نادرست به سؤال برای آزمودنی‌ای است که به همه مهارت‌های لازم برای پاسخ صحیح به سؤال تسلط یافته است. برای مثال سؤال ۱ نیازمند دو مهارت دستور زبان و استخراج اطلاعات صریح است. پارامتر حدس سؤال ۱ در گروه زنان ۰/۲۶ است، بنابراین آزمودنی که به دو مهارت لازم برای این سؤال مسلط نیست ۲۶ درصد شانس دارد پاسخ صحیح به سؤال بدهد. شاخص لغزش سؤال ۱ در گروه زنان ۰/۳۸ است. آزمودنی که به هر دو مهارت مسلط است، ۳۸ درصد احتمال

1. Loglikelihood
2. Akaike information criterion
3. Bayesian information criterion
4. deterministic inputs, noisy or gate
5. reduced reparametrized unified model
6. additive CDM
7. Guessing
8. slipping

وجود دارد که به سؤال پاسخ نادرست بدهد. مقادیر حدس برای گروه زنان بین ۰/۰۹ و ۰/۲۶ است؛ که می‌کهن حدس ۰/۲۰ است. مقادیر لغزش بین ۰/۱۴ و ۰/۸۵ قرار دارند. میانگین مقادیر لغزش برای گروه زنان ۰/۴۷ است. برای گروه مردان مقادیر حدس ۰/۰۴ تا ۰/۷۱ قرار دارند که دامنه وسیع‌تری نسبت گروه زنان دارد، میانگین مقادیر حدس ۰/۲۳ است. پارامترهای لغزش گروه مردان در دامنه ۰/۱۰ و ۰/۷۱ است و میانگین آن‌ها ۰/۳۷ است. مقایسه میانگین مقادیر حدس و لغزش آشکار می‌کند که در کل ممکن است سؤالات به ضرر زنان عمل کنند. به‌طور متوسط احتمال حدس سؤالات برای گروه مردان بیشتر از زنان است. این امر می‌تواند به این علت باشد که مقدار حدس برای سؤال‌های ۶ و ۲۱ در مردان بسیار بیشتر از زنان است. میانگین احتمال لغزش در سؤالات برای زنان بیشتر از مردان است. دو سؤال ۱۳ و ۲۳ سؤال‌هایی هستند که تفاوت لغزش زنان و مردان و بیشتر از سایر سؤال‌ها است. برخی از سؤالات مانند سؤال‌های شماره ۲، ۴، ۵، و ۱۰ مقادیر حدس و لغزش بسیار بالایی دارند، که به قدرت تمیز بسیار پایین این سؤالات می‌انجامد.

نتایج کنش افتراقی سؤال در جدول ۴ مشاهده می‌شود. برای کنترل نرخ خطای نوع اول از روش اصلاح سطح معناداری بنجامینی-هاچبرگ^۱ استفاده شد. آزمون والد ۱۳ سؤال را دارای کنش افتراقی سؤال تشخیص داده است که بیشترین تعداد را در بین سه روش به خود اختصاص داده است. آزمون نسبت درست‌نمایی اصلاح‌شده ۱۱ سؤال و آزمون نسبت درست‌نمایی ۵ سؤال را دارای کنش افتراقی سؤال تشخیص داده‌اند. ۱۶ سؤال مختلف توسط سه روش دارای کنش افتراقی سؤال تشخیص داده شدند که ۵۰٪ سؤالات را شامل می‌شود. سؤال‌های ۹، ۱۷، ۱۹، ۲۲ و ۲۳ توسط هر سه روش دارای کنش افتراقی سؤال تشخیص داده شده‌اند. با استفاده از ضریب توافق کاپا میزان توافق سه روش در تشخیص کنش افتراقی سؤال بررسی شد. ضریب توافق آزمون والد با آزمون نسبت درست‌نمایی ۰/۴۶، و با آزمون نسبت درست‌نمایی اصلاح‌شده ۰/۳۷ بود. ضریب توافق آزمون نسبت درست‌نمایی با آزمون نسبت درست‌نمایی اصلاح‌شده ۰/۵۱ بود. میانگین ضریب توافق سه روش ۰/۴۵ بود. ضرایب کاپا بین ۰/۴۱ و ۰/۷۰ به‌عنوان توافق متوسط تلقی می‌شوند (Altman, 1999)، بنابراین بین سه روش توافق متوسط وجود دارد.

سؤال‌های ۲، ۶، ۹، ۱۲، ۱۷، ۲۰، ۲۱، ۲۲، ۲۳، ۲۴ و ۲۵ دارای کنش افتراقی سؤال یکنواخت هستند. پارامتر حدس این سؤال‌ها در گروه مردان بیشتر از گروه زنان است و

1. Benjamini-Hochberg

پارامتر لغزش آن‌ها برای مردان کمتر از زنان است، از این رو این ۱۱ سؤال به نفع مردان و به ضرر زنان دارای کنش افتراقی سؤال هستند. برای مثال در سؤال شماره ۲، پارامتر حدس برای زنان ۰/۲۴ و برای مردان ۰/۲۷ است، بنابراین در بین داوطلبانی که به مهارت‌های مورد نیاز سؤال ۲ مسلط نیستند مردان در مقایسه با زنان شانس بالاتری برای پاسخ صحیح به سؤال دارند. پارامتر لغزش برای زنان و مردان به ترتیب ۰/۷۰ و ۰/۷۵ است، که نشان می‌دهد در بین آزمودنی‌هایی که به مهارت‌های لازم برای سؤال مسلط هستند، احتمال پاسخ نادرست برای زنان در مقایسه با مردان بیشتر است. کنش افتراقی یکنواخت در این سؤال به این معنا است که در هر دو حالت تسلط و عدم تسلط به مهارت‌ها مردان شانس بالاتری برای پاسخ درست به سؤال دارند.

سؤالات ۳، ۸، ۱۰، ۱۶ و ۱۹ دارای کنش افتراقی غیریکنواخت هستند. در این سؤالات هر دو پارامتر حدس و لغزش در زنان بیشتر از مردان است. در سؤال ۱۹، به عنوان نمونه، پارامتر حدس برای زنان ۰/۱۶ و برای مردان ۰/۰۶ است، بنابراین سؤال ۱۹ برای آزمودنی‌هایی که فاقد مهارت‌های لازم برای پاسخ صحیح به سؤال هستند به نفع زنان و به ضرر مردان عمل می‌کند. پارامتر لغزش سؤال ۱۹ برای زنان ۰/۶۱ و برای مردان ۰/۳۴ است، این امر نشان می‌دهد در بین آزمودنی‌هایی که همه مهارت‌های لازم برای پاسخ به سؤال را دارند مردان در مقایسه با زنان به احتمال بیشتری به سؤال پاسخ صحیح می‌دهند و سؤال در بین این گروه به نفع مردان دارای کنش افتراقی سؤال است.

جدول ۳. پارامتر حدس و پارامتر لغزش سؤالات به تفکیک جنسیت

سؤال	زنان		مردان	
	حدس	لغزش	حدس	لغزش
۱	۰/۲۶	۰/۳۸	۰/۱۴	۰/۲۴
۲	۰/۲۴	۰/۷۵	۰/۲۷	۰/۷۰
۳	۰/۲۱	۰/۳۶	۰/۱۹	۰/۳۰
۴	۰/۱۸	۰/۷۳	۰/۲۸	۰/۷۱
۵	۰/۲۶	۰/۵۱	۰/۲۶	۰/۵۳
۶	۰/۱۹	۰/۴۳	۰/۵۷	۰/۴۱
۷	۰/۲۶	۰/۲۴	۰/۲۷	۰/۱۹
۸	۰/۲۳	۰/۵۳	۰/۰۱	۰/۴۱
۹	۰/۱۸	۰/۶۸	۰/۲۳	۰/۵۵

سؤال	زنان		مردان	
	حدس	لغزش	حدس	لغزش
۱۰	۰/۱۵	۰/۸۵	۰/۱۷	۰/۷۰
۱۱	۰/۱۷	۰/۶۵	۰/۰۴	۰/۵۵
۱۲	۰/۱۷	۰/۶۴	۰/۱۹	۰/۵۹
۱۳	۰/۱۹	۰/۶۱	۰/۲۳	۰/۱۲
۱۴	۰/۲۶	۰/۲۹	۰/۲۶	۰/۳۳
۱۵	۰/۲۴	۰/۲۹	۰/۲۴	۰/۳۵
۱۶	۰/۱۹	۰/۶۹	۰/۰۲	۰/۴۶
۱۷	۰/۱۷	۰/۵۲	۰/۲۵	۰/۳۵
۱۸	۰/۱۹	۰/۶۵	۰/۲۰	۰/۵۷
۱۹	۰/۱۶	۰/۶۱	۰/۰۶	۰/۳۴
۲۰	۰/۱۵	۰/۴۸	۰/۱۹	۰/۳۱
۲۱	۰/۱۶	۰/۳۴	۰/۷۱	۰/۱۸
۲۲	۰/۰۹	۰/۳۳	۰/۱۱	۰/۱۰
۲۳	۰/۲۱	۰/۴۴	۰/۲۴	۰/۱۹
۲۴	۰/۱۸	۰/۲۴	۰/۲۲	۰/۱۲
۲۵	۰/۲۱	۰/۲۷	۰/۲۴	۰/۱۶
۲۶	۰/۲۵	۰/۳۸	۰/۱۵	۰/۴۹
۲۷	۰/۱۶	۰/۳۵	۰/۱۸	۰/۲۹
۲۸	۰/۱۵	۰/۴۳	۰/۱۶	۰/۳۵
۲۹	۰/۲۶	۰/۱۴	۰/۲۶	۰/۱۴
۳۰	۰/۱۹	۰/۴۹	۰/۲۰	۰/۵۶

جدول ۴. نتایج آزمون والد، نسبت درستنمایی و نسبت درستنمایی اصلاح شده برای ارزیابی کنش افتراقی سؤال

سؤال	درجه آزادی	CDM-WALD		CDM-LR		R-CDM-LR	
		آماره	سطح معناداری #	آماره	سطح معناداری #	آماره	سطح معناداری #
۱	۴	۵/۵۰	۰/۳۱۳	۹/۷۰	۰/۱۵۲	-۱۳/۴۲	۱/۰۰۰
۲	۲	۶/۷۷	۰/۰۷۲	۸/۶۸	۰/۰۶۵	۱۰/۸۲	۰/۰۱۴
۳	۴	۱۴/۶۹	۰/۰۱۲	۳/۸۵	۰/۶۳۶	۱۲/۴۷	۰/۰۴۲
۴	۸	۸/۳۸	۰/۴۷۶	۵/۰۲	۰/۸۳۴	۱۲/۰۷	۰/۲۴۶
۵	۴	۴/۶۸	۰/۴۰۱	۳/۱۱	۰/۶۷۳	۶/۸۷	۰/۲۴۶
۶	۴	۴۳/۵۷	۰/۰۰۱	۱/۳۱	۰/۸۵۹	-۴۹/۴۸	۱/۰۰۰

سؤال	درجه آزادی	CDM-WALD		CDM-LR		R-CDM-LR	
		آماره	سطح معناداری #	آماره	سطح معناداری #	آماره	سطح معناداری #
۷	۲	۴/۹۱	۰/۱۵۰	۰/۰۳۸	۰/۸۵۲	-۱۰/۹۰	۱/۰۰۰
۸	۸	۲۸/۰۶	۰/۰۰۲	۱۶/۵۱	۰/۱۳۳	۲۸/۸۱	۰/۰۰۲
۹	۲	۱۱/۹۹	۰/۰۰۶	۱۲/۲۴	۰/۰۱۹	۲۴/۳۶	۰/۰۰۱
۱۰	۴	۱۰/۰۳	۰/۰۷۹	۹/۱۳	۰/۱۷۳	۲۴/۱۹	۰/۰۰۱
۱۱	۴	۱۲/۱۲	۰/۷۶۳	۶/۴۰	۰/۳۴۸	۴/۷۲	۰/۴۳۲
۱۲	۴	۸/۲۹	۰/۱۵۰	۲/۹۲	۰/۶۸۳	۱۲/۱۵	۰/۰۴۴
۱۳	۴	۷/۳۸	۰/۱۹۵	۶/۳۵	۰/۳۴۸	۸/۵۳	۰/۱۷۰
۱۴	۴	۵/۴۹	۰/۳۱۳	۴/۸۰	۰/۵۰۸	۸/۲۸	۰/۱۷۴
۱۵	۲	۰/۱۳	۰/۹۳۳	۴/۱۷	۰/۲۸۶	۰/۱۵	۱/۰۰۰
۱۶	۸	۲۵/۴۰	۰/۰۰۴	۱۸/۳۴	۰/۰۸۰	۲۴/۷۵	۰/۰۰۶
۱۷	۴	۳۷/۸۹	۰/۰۰۱	۱۹/۰۰	۰/۰۱۱	۳۴/۱۱	۰/۰۰۱
۱۸	۴	۵/۵۹	۰/۳۱۳	۱/۷۶	۰/۸۳۴	۵/۶۸	۰/۳۳۶
۱۹	۴	۵۰/۵۰	۰/۰۰۱	۱۶/۲۸	۰/۰۱۹	۲۵/۵۰	۰/۰۰۱
۲۰	۴	۱۹/۲۷	۰/۰۰۳	۸/۵۸	۰/۱۸۰	۵/۶۱	۰/۳۳۶
۲۱	۸	۲۵/۷۵	۰/۰۰۴	۷/۷۶	۰/۶۳۶	۸/۸۳	۰/۴۶۴
۲۲	۲	۱۴/۱۰	۰/۰۰۳	۱۱/۴۸	۰/۰۱۹	۱۳/۸۱	۰/۰۰۴
۲۳	۲	۲۷/۶۳	۰/۰۰۱	۱۷/۵۹	۰/۰۰۴	۳۰/۹۹	۰/۰۰۱
۲۴	۴	۱۵/۲۳	۰/۰۱۱	۵/۱۳	۰/۴۸۸	۱۰/۷۸	۰/۰۷۲
۲۵	۴	۲۲/۱۱	۰/۰۰۱	۳/۵۷	۰/۶۳۶	-۵/۳۵	۱/۰۰۰
۲۶	۴	۳/۹۳	۰/۴۷۸	۸/۷۸	۰/۱۸۰	۵/۵۵	۰/۳۳۶
۲۷	۲	۳/۲۵	۰/۲۹۳	۰/۵۵	۰/۸۳۴	۳/۸۸	۰/۲۴۶
۲۸	۲	۳/۹۸	۰/۲۱۵	۲/۲۶	۰/۵۰۸	۳/۹۳	۰/۲۴۶
۲۹	۴	۲/۸۵	۰/۶۴۷	۵/۱۰	۰/۴۸۸	۳/۳۰	۰/۶۳۵
۳۰	۴	۱/۶۵	۰/۸۲۶	۳/۳۸	۰/۶۴۶	۲/۷۴	۰/۷۲۲

#سطح معناداری اصلاح شده به روش بنجامینی-هاچبرگ

بحث و نتیجه گیری

مدل‌های تشخیصی شناختی توجه فزاینده‌ای را در سال‌های اخیر به خود جلب کرده‌اند. این تحقیق به منظور ارزیابی کنش افتراقی سؤال نسبت به جنسیت برای سؤالات درس زبان انگلیسی آزمون کارشناسی ارشد مجموعه روان‌شناسی سال ۱۳۹۶ با استفاده از شیوه‌های موجود در بافت مدل‌های تشخیصی شناختی طراحی و اجرا گردید. ماتریس Q بر اساس

مهارت‌های واژگان، دستور زبان، استخراج اطلاعات صریح، اتصال و ترکیب کردن، و استنتاج طراحی و اعتبار آن تأیید شد. این یافته با تحقیقاتی که این پنج مهارت را به کار برده اند همخوان است (برای مثال، Jang, 2009؛ Ravand et al., 2012؛ Li, 2011). بررسی پارامترهای سؤالات حاکی از بالا بودن پارامترهای حدس و مخصوصاً لغزش سؤالات آزمون زبان انگلیسی بود که نشان‌دهنده کیفیت پایین سؤال‌ها در قالب مدل تشخیصی شناختی است. در ارتباط با سؤال اول تحقیق، بر اساس یافته‌ها آزمون نسبت درستنمایی محافظه‌کارترین روش تشخیص کنش افتراقی سؤال بود، هر ۵ سؤالی که آزمون نسبت درستنمایی به‌عنوان دارای کنش افتراقی سؤال مشخص کرده توسط دو روش دیگر نیز دارای کنش افتراقی شناسایی شده‌اند. این یافته با توجه به تعداد بالای سؤالات دارای کنش افتراقی با مطالعه لئو و همکاران (Liu et al., 2019) همخوان است که با افزایش تعداد سؤالات دارای کنش افتراقی سؤال توان آزمون نسبت درستنمایی کاهش می‌یابد. روش نسبت درستنمایی اصلاح شده ۱۱ و آزمون والد ۱۳ سؤال را شناسایی کردند. تحقیقات قبلی (Ma et al., 2017؛ Ma et al., 2021) نشان داده آزمون نسبت درستنمایی اصلاح شده در تحقیقات شبیه‌سازی در بیشتر شرایط عملکرد قابل‌اعتمادی داشته است، این روش در شرایط کیفیت پایین سؤالات در حجم نمونه‌های متوسط و بزرگ از نظر خطای نوع اول بهترین عملکرد را در بین سه روش داشته است؛ بنابراین با توجه به کیفیت نسبتاً پایین سؤالات و حجم نمونه بالا می‌توان گفت نتایج آزمون نسبت درستنمایی اصلاح شده قابل‌اعتمادتر هستند. آزمون والد بیشترین سؤال را دارای کنش افتراقی شناسایی کرد. نکته‌ای که در این ارتباط باید در نظر داشت این است که بر اساس تحقیقات پیشین (Hou, 2013؛ Hou et al., 2014)، روش والد در مدل G-DINA با افزایش پارامترهای حدس و لغزش نرخ‌های خطای نوع اول بیشتر از مقدار اسمی دارد. با توجه به پارامترهای حدس و لغزش سؤالات آزمون زبان انگلیسی، می‌توان تعداد بالای سؤالات شناسایی شده توسط آزمون والد را به کیفیت سؤالات مرتبط دانست.

ضریب توافق کاپا حاکی توافق متوسط سه روش در شناسایی کنش افتراقی سؤال بود. بیشترین توافق بین نسبت درستنمایی و نسبت درستنمایی اصلاح شده مشاهده شد. این یافته به علت این که این دو روش بر یک آماره متکی هستند، قابل‌انتظار بود. آزمون والد با نسبت درستنمایی اصلاح شده توافق بالاتری داشت. این امر را می‌توان به شباهت فرایند اجرای

آزمون نسبت درستنمایی اصلاح شده با آزمون والد نسبت داد که کمتر تحت تأثیر آلودگی برآورد پروفایل ها قرار می گیرد. در روش والد یک مدل دو گروهی بدون محدودیت بر روی پارامترها اجرا می شود که پارامترهای همه سؤالات به صورت آزادانه در هر دو گروه برآورد می شوند. در آزمون نسبت درستنمایی اصلاح شده در هر بار فقط پارامترهای یک سؤال در دو گروه یکسان فرض می شود و سایر سؤالات به صورت آزاد در دو گروه برآورد می شوند. عدم محدودیت در آزمون والد و وجود یک محدودیت در نسبت درستنمایی اصلاح شده موجب می شود تا پروفایل های مکنون برآورد شده توسط دو روش در مقایسه با آزمون نسبت درستنمایی کمتر آلوده باشند؛ زیرا در آزمون نسبت درستنمایی در هر بار فقط پارامترهای یک سؤال به صورت آزادانه برآورد می شود و سایر سؤالات در دو گروه محدود هستند و نقش لنگر را دارند. در مجموع آزمون نسبت درستنمایی اصلاح شده روش کارآمدی برای شناسایی کنش افتراقی سؤال در مدل های تشخیصی شناختی است و نتایج آن نزدیک به آزمون والد است.

در ارتباط با سؤال دوم تحقیق، در مجموع ۱۶ سؤال دارای کنش افتراقی سؤال نسبت به جنسیت بودند. از این تعداد، ۱۱ سؤال دارای کنش افتراقی سؤال یکنواخت به ضرر زنان بودند. این یافته با تحقیقات قبلی بر روی درک مطلب آزمون زبان انگلیسی که کنش افتراقی سؤال به ضرر زنان را نشان داده اند همخوان است (Amirian et al., 2014؛ Mehrzmay et al., 2021؛ Jamalzadeh et al., 2022). از مجموع ۱۶ سؤال دارای کنش افتراقی، مهارت واژگان در ۱۱ سؤال، دستور زبان در ۶ سؤال، اتصال و ترکیب در ۹ سؤال، استنتاج در ۳ سؤال، و استخراج اطلاعات صریح در ۱ سؤال نقش داشتند. سؤالات شماره ۱۹-۲۴ مربوط به متن شماره ۴ بودند که همه آنها دارای کنش افتراقی بودند که از بین آنها ۵ سؤال کنش افتراقی به نفع مردان داشتند. محتوای این متن می تواند منبع احتمالی کنش افتراقی باشد. این متن در مورد استفاده از پرسشنامه های خود گزارشی و روش های مشاهده ای در ارزیابی روانی بیماران مبتلا به سکته بود. مطالعات قبلی نشان داده است که مردان در سؤالاتی که مرتبط با علم و محتواهای فنی است عملکرد بهتری دارند (Doolittle et al., 1988؛ Gafni, 1991). وجود کنش افتراقی سؤال به دلیل لطمه به روایی آزمون، هرگونه تصمیم گیری در مورد پذیرش دانشجویان را عمدتاً در معرض خطا و اجحاف در حق شرکت کنندگان زن در کنکور کارشناسی ارشد مجموعه روان شناسی قرار می دهد (بنی اسدی، ۱۴۰۱).

یکی از محدودیت‌های این پژوهش این بود که مقایسه سه روش ارزیابی کنش افتراقی سؤال بر روی داده‌های واقعی صورت گرفت، در این شرایط امکان قضاوت در مورد دقت روش‌ها امکان‌پذیر نیست و تنها توافق آن‌ها را می‌توان موردبررسی قرار گرفت. با توجه به تفاوت سه روش در سؤالات دارای کنش افتراقی سؤال، انجام مطالعات شبیه‌سازی با استفاده از پارامترهای واقعی سؤال، قضاوت دقیق‌تر در مورد کارایی روش‌ها در شرایط مختلف را امکان‌پذیر می‌کند. در تحقیق حاضر کنش افتراقی سؤال تنها نسبت به جنسیت موردبررسی قرار گرفت، متغیرهای جمعیت شناختی دیگر مانند زبان مادری، مجموعه مقطع کارشناسی، و منطقه سکونت داوطلب می‌توانند منابع احتمالی دیگر برای کنش افتراقی سؤال باشند که موردبررسی قرار نگرفته‌اند.

هرچند وجود کنش افتراقی سؤال به معنای سوگیری سؤال نیست. انجام مطالعات بیشتر به‌منظور شناسایی منبع واریانس غیرمرتبط با سازه که منجر به کنش افتراقی سؤال شده است ضروری می‌نماید. با توجه به حساسیت و مسائل امنیتی که امکان ارزیابی کنش افتراقی سؤال پیش از اجرای نهایی کنکور را فراهم نمی‌کند، تشکیل کمیته‌هایی متشکل از متخصصان موضوع برای بررسی محتوای سؤالات پیش از اجرای آزمون به‌منظور قضاوت در مورد وجود سوگیری احتمالی می‌تواند به رعایت بیشتر عادلانه بودن آزمون کمک کند.

منابع

- بنی اسدی، علی، صالحی، کیوان، خدایی، ابراهیم، باقری، خسرو و ایزانلو، بلال. (۱۴۰۱). رویکرد سنجش کلاسی عادلانه: برازش مدل چندبعدی پاسخ مدرج. *فصلنامه اندازه‌گیری تربیتی*، ۱۳(۴۹)، ۵۱-۳۱. doi: 10.22054/jem.2023.63328.2219
- کاظمی دانا، بهروز. (۱۳۹۵). کارکرد افتراقی سؤالات‌های (DIF) آزمون ریاضی پایه هشتم تیمز ۲۰۰۷ در بین دانش‌آموزان دختر و پسر با استفاده از مدل‌های تشخیصی شناختی (CDMs). پایان‌نامه کارشناسی ارشد. دانشگاه علامه طباطبائی.
- سعادت، سمیه، مقدم‌زاده، علی، مینایی، اصغر، و گرامی‌پور، مسعود. (۱۳۹۹). کارکرد افتراقی سؤال در چارچوب سنجش تشخیصی شناختی: مطالعه موردی سؤال‌های حساب دیفرانسیل و انتگرال کنکور سراسری ۹۷. *دو فصلنامه راهبردهای شناختی در یادگیری*، ۸(۱۵)، ۱۹-۳۵.

References

- Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, 29(1), 67-91. <https://doi.org/10.1111/j.1745-3984.1992.tb00370.x>
- Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing*, 22(3), 301-320. <https://doi.org/10.1191/0265532205lt310oa>
- Altman, D. G. (1999). *Practical statistics for medical research*. Chapman & Hall/CRC Press.
- Amirian, S. M. R., Alavi, S. M., & Fidalgo, A. M. (2014). Detecting gender DIF with an English proficiency test in EFL context. *Iranian Journal of Language Testing*, 4(2), 187-203.
- Baniasadi, A., Salehi, K., Khodaie, E., Bagheri, K., & Izanloo, B. (2022). Fair Classroom Assessment Rubric: Fitting a Multidimensional Graded Response Model. *Educational Measurement*, 13(49), 31-51. [In Persian]
- de la Torre, J., & Chiu, C. Y. (2016). A general method of empirical Q-matrix validation. *Psychometrika*, 81(2), 253-273. <https://doi.org/10.1007/s11336-015-9467-8>
- Doolittle, A., & Welch, C. (1989). Gender differences in performance on a college-level achievement test. *Academic Achievement*, 12(3), 45-62.
- Effatpanah, F., Baghaei, P., & Boori, A. A. (2019). Diagnosing EFL learners' writing ability: A diagnostic classification modeling analysis. *Language Testing in Asia*, 9(1), 12. <https://doi.org/10.1186/s40468-019-0086-7>
- Gafni, N. (1991). Differential item functioning: Performance by sex on reading comprehension tests. *Journal of Educational Assessment*, 8(2), 123-145.
- George, A. C., & Robitzsch, A. (2014). Multiple group cognitive diagnosis models, with an emphasis on differential item functioning. *Psychological Test and Assessment Modeling*, 56(4), 405-432.

- Hou, L. (2013). *Differential item functioning assessment in cognitive diagnostic modeling: Applying the Wald test to investigate DIF in the generalized DINA model framework* [Doctoral dissertation, Rutgers University]. ProQuest Dissertations Publishing.
- Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, 51(1), 98-125. <https://doi.org/10.1111/jedm.12034>
- Hou, L., Terzi, R., & de la Torre, J. (2021). Wald test formulations in DIF detection of CDM data with the proportional reasoning test. *International Journal of Assessment Tools in Education*, 7(2), 145-158. <https://doi.org/10.21449/ijate.706425>
- Hunter, C. V., Li, H., & Liu, R. (2022). Methods to retrofit and validate Q-matrices for cognitive diagnostic modeling. In M. Wiberg (Ed.), *Quantitative psychology* (pp. 217-225). Springer. https://doi.org/10.1007/978-3-030-81272-3_17
- Jamalzadeh, M., Lotfi, A., & Rostami, M. (2022). Equity on general English achievement tests through gender-based DIF analysis across different majors. *International Journal of Foreign Language Teaching and Research*, 10(4), 47-65.
- Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for applying Fusion Model to LanguEdge assessment. *Language Testing*, 26(1), 31-73. <https://doi.org/10.1177/0265532208097336>
- Kazemi Dana, B. (2016). Examining Differential Item Functioning (DIF) of the 2007 TIMSS 8th grade mathematics test among male and female students using of Cognitive Diagnostic Models (CDMs) [Unpublished master's dissertation]. Allameh Tabataba'i University. [In Persian]
- Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.
- Li, H., Hunter, C. V., & Lei, P. W. (2015). The selection of cognitive diagnostic models for a reading comprehension test. *Language Testing*, 33(3), 391-409. <https://doi.org/10.1177/0265532215590848>
- Li, X., & Wang, W. C. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement*, 52(1), 28-54. <https://doi.org/10.1111/jedm.12060>
- Liu, Y., Yin, H., Xin, T., Shao, L., & Yuan, L. (2019). A comparison of differential item functioning detection methods in cognitive diagnostic models. *Frontiers in Psychology*, 10, 1073. <https://doi.org/10.3389/fpsyg.2019.01073>
- Ma, W. (2018). *GDINA: The generalized DINA model framework* (Version 2.1) [Computer software]. <https://CRAN.R-project.org/package=GDINA>
- Ma, W., Terzi, R., & de la Torre, J. (2021). Detecting differential item functioning using multiple-group cognitive diagnosis models. *Applied Psychological Measurement*, 45(1), 37-53. <https://doi.org/10.1177/0146621620959829>
- Ma, W., Terzi, R., Lee, S., & de la Torre, J. (2017). Multiple-group cognitive diagnosis models and their applications in detecting differential item functioning. *Educational Measurement: Issues and Practice*, 36(3), 28-40.
- Mehrazmay, R., Ghonsooly, B., & de la Torre, J. (2021). Detecting differential item functioning using cognitive diagnosis models: Applications of the Wald test and likelihood ratio test in a university entrance examination. *Applied Measurement in Education*, 34(4), 262-284. <https://doi.org/10.1080/08957347.2021.1933981>

- Milewski, G. B., & Baron, P. A. (2002). Extending DIF methods to inform aggregate report on cognitive skills. *Educational Assessment*, 8(2), 123-145.
- Paulsen, J., Svetina, D., Feng, Y., & Valdivia, M. (2020). Examining the impact of differential item functioning on classification accuracy in cognitive diagnostic models. *Applied Psychological Measurement*, 44(4), 267-281. <https://doi.org/10.1177/0146621619889098>
- Qiu-Yun, L. I., Yan, S. A., Tai-Xun, W. A. N. G., & Dong-Bei, C. H. U. (2022). DIF detection methods and its application of polytomously scored items under the framework of cognitive diagnosis. *Journal of Psychological Science*, 45(4), 998.
- Ravand, H., Barati, H., & Widhiarso, W. (2012). Exploring diagnostic capacity of a high stakes reading comprehension test: A pedagogical demonstration. *Iranian Journal of Language Testing*, 3(1), 12-37.
- Revelle, W. (2018). *psych: Procedures for personality and psychological research* (Version 1.8.10) [Computer software]. Northwestern University. <https://CRAN.R-project.org/package=psych>
- Saadati, S., Moghadamzadeh, A., Minaei, A., & Geramipour, M. (2020). Differential item functioning in the framework of cognitive diagnostic assessment: Questions related to the differential and integral calculus of the Iranian national university entrance examination 2018. *Biquarterly Journal of Cognitive Strategies in Learning*, 8(15), 19-35. [In Persian]
- Svetina, D., Feng, Y., Paulsen, J., Valdivia, M., Valdivia, A., & Dai, S. (2018). Examining DIF in the context of CDMs when the Q-matrix is misspecified. *Frontiers in Psychology*, 9, 2365. <https://doi.org/10.3389/fpsyg.2018.02365>
- Urmston, A., Raquel, M., & Tsang, C. (2013). Diagnostic testing of Hong Kong tertiary students' English language proficiency: The development and validation of DELTA. *Hong Kong Journal of Applied Linguistics*, 14(2), 60-82.
- Wang, C., & Gierl, M. J. (2011). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in critical reading. *Journal of Educational Measurement*, 48(2), 165-187. <https://doi.org/10.1111/j.1745-3984.2011.00142.x>
- Wang, N., & Wang, C. (2022). Exploring differential item functioning in cognitive diagnostic computer adaptive testing: A simulation study. *Educational and Psychological Measurement*, 82(1), 121-145.
- Yi, Y. (2012). *Implementing a cognitive diagnostic assessment in an institutional test: A new networking model in language testing and experiment with a new psychometric model and task type* [Doctoral dissertation, University of Illinois at Urbana-Champaign].
- Zhang, W. (2006). *Detecting differential item functioning using the DINA model* [Doctoral dissertation, University of Maryland].