فصلنامهٔ روان‌شناسی کاربردی
سال۴، شماره ۲(۱۴)، تابستان ۱۳۸۹، ۳۱ - ۲۰

Journal of Applied Psychology
Vol. 4, No. 2(14), Summer 2010, 20-31

شناسایی کارکرد متفاوت سؤال براساس نظریهٔ سؤال- پاسخ: کاربرد الگوی تک-پارامتری با استفاده از نرم‌افزار بای‌لوگ-ام‌جی

Identifying differential item function (DIF) based on item-response theory: Application of the one parameter model using the BILOG-MG software

بلال ایزانلو: دانشجوی دکتری سنجش آموزش دانشگاه تهران
مجتبی حبیبی عسگرآبادی: دانشجوی دکترای روان‌شناسی سلامت دانشگاه تهران، بورسیه هیأت علمی پژوهشکده خانواده دانشگاه شهید بهشتی

چکیده

هدف: مقالهٔ حاضر به دنبال توصیف و تشریح روش تشخیص کارکرد متفاوت سؤال در چارچوب نظریهٔ سؤال-پاسخ، با استفاده از الگوی یک پارامتری با تأکید بر سوگیری داخلی در یک آزمون است. **روش:** ابتدا مبانی نظری رویکرد معرفی شد، و به دنبال آن به منظور ارائه عملی روش، داده‌های خرده‌مقیاس گرامر در بخش تخصصی دفترچه زبان انگلیسی آزمون سراسری در سال ۱۳۸۴ با نرم‌افزار بای‌لوگ- ام‌جی تحت الگوی یک پارامتری با دو روش مقایسهٔ سطوح مقادیر دشواری در دو گروه و آزمون نسبت درست‌نمایی تحلیل شدند. **یافته‌ها**: نتایج نشان داد که بین مقدار دشواری سؤال ۸ در دو گروه تفاوت معنادار است. این موضوع با برازش بهتر الگوی دارای کارکرد متفاوت سؤال در مقایسه با الگوی بدون کارکرد متفاوت سؤال (با آلفای ۰/۰۵) تأیید شد. در واقع احتمال پاسخ درست به سؤال ۸ برای گروه مردان در مقایسه با گروه زنان بیشتر بوده است. **نتیجه-گیری:** اهمیت بررسی کارکرد متفاوت سؤال در آزمون‌های مختلف روز به روز در حال افزایش است. روش‌های موجود در نظریهٔ کلاسیک به خاطر وجود تغییرپذیری شاخص‌های آن برای شناسایی این موضوع مفید نیستند. در مقابل روش‌هایی موجو در بافت نظریه سؤال-پاسخ برای انجام این کار مفیدتر هستند.

کلیدواژه‌ها: *الگوی تک‌پارامتری، کارکرد متفاوت سؤال، نرم افزار بای‌لوگ-ام جی، نظریه سؤال- پاسخ*

B. Ezanloo: PhD student in Educational Measurement, Tehran Uni.
E-mail: *b.ezanloo@gmail.com*
M. Habibi Asgarabadi: PhD student of Health Psychology of Tehran Uni. & Family Research Institute, Shahid Beheshti Uni, G. C.

Abstract

Aim: The aim of the present paper was to describe and explain the method of differential item function (DIF) within the framework of the item response theory (IRT) using the one parameter model with an emphasis on internal answer bias to the test. **Method:** As a first step, the theoretical foundations of the perspective was introduced and then in order to offer a practical application of the method, items of the English grammar subscale of the foreign language booklet in the National University Entrance Exam (NUEM) in the academic year 2005–06 was analyzed by two methods of comparing levels of difficulty values in two groups and likelihood ratio test. This was done by using BILOG-MG software under the one parameter model. **Results:** Findings showed there was a significant difference in the amount of difficulty coefficient between the two groups on question number 8. This was confirmed by the fact that there was a better fitness when using the model including DIF compared with the model without DIF ($p<0.05$). In fact the probability of a correct response was higher for males than for females. **Conclusion:** The importance of investigating DIF in different tests is becoming increasingly important The methods for investigating DIF in classical test theory are not useful due to the variability in the indices obtained by this method. In contrast, techniques under the IRT context seem more plausible.

Keywords: *BILOG-MG software, differential item functions, item response theory, one parameter model*

دریافت مقاله: ۱۳۸۸/۹/۱۶؛ **دریافت نسخه نهایی**: ۱۳۸۹/۷/۲۶؛ **پذیرش مقاله**: ۱۳۸۹/۸/۱۰

مقدمه

کاربرد گستردهٔ آزمون‌ها و استفاده از نتایج آن‌ها در تصمیم‌گیری و طبقه‌بندی افراد، که روز به روز در حال گسترش است، به طرح این سؤال می‌انجامد که آیا امکان دارد سؤال‌های موجود یک آزمون یا خود یک آزمون تحت تأثیر متغیرهایی مثل جنس، اقلیت، زبان مادری و جز آن کارکرد متفاوتی داشته باشد؟ این موضوع که در نظریه‌های کلاسیک و سؤال-پاسخ به ترتیب به سوگیری[1] و کارکرد متفاوت سؤال[2] معروف است و طیف بسیار وسیعی از پژوهش‌های نظری و عملی را به خود اختصاص داده است (امبرستون و رایس[3]، ۲۰۰۰).

سوگیری آزمون یا سؤال به بحث روایی[4] مربوط است؛ زیرا سوگیری در یک سؤال یا آزمون حاکی از وجود عامل مؤثر دیگری در پاسخ‌گویی به یک سؤال یا آزمون است که به هیچ وجه سازندهٔ آزمون به دنبال اندازه‌گیری آن نیست (راجو، لافیت و برن[5]، ۲۰۰۲). بنابراین، سؤالاتی که چنین کارکردی داشته باشند، باید تعدیل، اصلاح یا حذف شود، در غیر این صورت منجر به برآورد نادرست پارامترهای سؤالات و افراد می‌شود، موضوعی که برک[6] (۱۹۸۲) نیز به آن اشاره کرده و می‌تواند عواقب خطرناکی در نتایج آزمون داشته باشد.

اجزاء یک آزمون سؤال‌های موجود در آن هستند، پس برای آن‌که آزمونی سوگیری نداشته باشد، باید اول سوگیری سؤال‌های آن را بررسی کرد. البته نبود سؤال‌های دارای سوگیری در یک آزمون، شرط لازم برای رسیدن به یک آزمون فاقد سوگیری است، ولی شرط کافی نیست. اگر سؤالات آزمون یا عوامل دیگری باعث شود که آزمون نتواند هدف مورد اندازه‌گیری را منعکس کند مفهوم سنجش عادلانه به مثابه یکی از اهداف سنجش و اندازه‌گیری توانایی‌ها و ویژگی‌های افراد نقض می‌شود و کلیهٔ تصمیم‌های مبتنی بر آزمون نادرست خواهد بود (فتوحی، ۱۳۸۷).

در بحث سوگیری ابزار، باید بین دو نوع سوگیری تفاوت قائل شد: سوگیری خارجی و داخلی (دراسگو[7]، ۱۹۸۲، ۱۹۸۷، به نقل از امبرستون و رایس، ۲۰۰۰). سوگیری خارجی زمانی روی می‌دهد که نمره‌های آزمون دو یا چند گروه از آزمودنی‌های موجود در جامعه با سایر متغیرهای آزمون نشده همبستگی‌های متفاوت داشته باشد. این موضوع به روایی پیش‌بین متفاوت ابزار اندازه‌گیری منجر می‌شود. این نوع سوگیری ممکن است بر حسب بافت یا شرایط کاربرد آزمون،

1. bias
2. differential item function(DIF)
3. Embereston & Reise
4. validity
5. Raju & Laffitte & Byrne
6. Berk
7. Drasgow

مهم و یا بی‌اهمیت باشد. سوگیری داخلی زمانی رخ می‌دهد که کوواریانس بین سؤال‌های یک آزمون برای دو یا چند گروه از آزمون‌دهندگان متفاوت باشد، که به آن سوگیری اندازه‌گیری گفته و سبب می‌شود که مقیاس اندازه‌گیری در بین گروه‌ها یکسان نباشد.

تقسیم‌بندی دیگر در این خصوص، سوگیری یکنواخت و غیریکنواخت است. براساس گفتهٔ ملنبرج[1] (۱۹۸۲، به نقل از نارایانان و سوامیناتان[2]، ۱۹۹۶) در داده‌های دومقوله‌ای آموزشی دو نوع کارکرد متفاوت سؤال ممکن است روی دهد. سوگیری یکنواخت زمانی رخ می‌دهد که بین سطح توانایی و عضویت در گروه، تعاملی وجود نداشته باشد؛ و سوگیری غیریکنواخت نیز زمانی رخ می‌دهد که بین سطح توانایی و عضویت در گروه، تعامل وجود داشته باشد. با این‌که سوگیری یکنواخت در آزمون‌های استاندارد بیشتر از سوگیری غیریکنواخت روی می‌دهد، ولی سؤال‌هایی که کارکردی غیریکنواخت دارند در داده‌های واقعی بیشتر مشخص شده‌اند.

جدول۱. روش‌های استفاده شده در تشخیص کارکرد متفاوت سؤال (به نقل از سایرسی و آلالوف[3]، ۲۰۰۳)

روش	منابع	کاربرد مناسب
نمودار دلتا	آنگوف، ۱۹۸۲؛ آنگوف و فورد[4]،۱۹۷۳	داده‌های دومقوله‌ای
استاندارد سازی	دورانز و کالیک[5]، ۱۹۸۶؛ دورانز و هالند[6]،	داده‌های دومقوله‌ای
منتل- هنزل	هالند و ثایر[7]، ۱۹۸۸؛ دورانز و هالند، ۱۹۹۳	داده‌های دومقوله‌ای
رگرسیون لوجستیک	سوامیناتان و روگرز[8]، ۱۹۹۰؛ کلاسر، نانگستر، مازور و ریپکی[9]، ۱۹۹۶	داده‌های دومقوله‌ای، داده‌های چندمقوله‌ای، همتاسازی چندمتغیری[10]
کای دو لرد[11]	لرد،۱۹۸۰	داده‌های دومقوله‌ای
مساحت بین منحنی‌ها در نظریهٔ سؤال-پاسخ[12]	راجو[13]، ۱۹۸۸	داده‌های دومقوله‌ای، داده‌های چندمقوله‌ای
نسبت درست‌نمایی در نظریهٔ سؤال-پاسخ[14]	تیسن، استینبرگ و وایمر[15]، ۱۹۸۸، ۱۹۹۳	داده‌های دومقوله‌ای، داده‌های چندمقوله‌ای
آزمون سوگیری شبیه سازی سؤال[1]	شلی و استوت[2]، ۱۹۹۳.	داده‌های دومقوله‌ای

1. Mellenbergh
2. Narayanan, & Swaminathan
3. Sireci & Allalouf
4. Ford
5. Dorans and Kulick
6. Holland
7. Thayer
8. Swaminathan and Rogers
9. Clauser, Nungester, Mazor and Ripkey
10. multivariate matching
11. Lord’s chi-square
12. IRT area
13. Raju
14. IRT likelihood ratio
15. Thissen, Steinberg & Wainer

جدول۱ پرکاربردترین روش‌هایی را که برای بررسی کارکرد متفاوت سؤال مورد استفاده قرار گرفته نشان می‌دهد (سایرسی و آلالوف، ۲۰۰۳). در چارچوب نظریهٔ کلاسیک، این روش‌ها عبارت‌اند از: ۱) نمودار دلتا که توسط آنگوف[3] در ۱۹۷۲ مطرح شد؛ ۲) روش کای‌دو کامل[4] که توسط شونمان[5] در ۱۹۷۹ ارائه گردید؛ ۳) روش منتل هنزل[6] که نوعی کای‌دو و بسطی از روش شونمان است (فتوحی، ۱۳۸۷) و ۴) مقایسهٔ پارامترهای مختلف سؤال با روش‌های متداول آماری در دو یا چند گروه (امبرستون و رایس، ۲۰۰۰). روش‌های رگرسیون لوجستیک و لگاریتم خطی نیز برای این منظور استفاده شده‌اند (میلساپ و اورسون[7]، ۱۹۹۳). در نظریهٔ خصیصه مکنون روش‌هایی مثل مقایسهٔ پارامترهای برآوردشده با الگوی یک پارامتری، مقایسهٔ منحنی‌های ویژگی‌های سؤال برای گروه‌های مختلف (معمولاً منحنی ویژگی سؤال الگوی ۳ پارامتری مورد استفاده قرار می‌گیرد) و روش‌های مبتنی بر آزمون برازش الگو برای این منظور استفاده شد (فتوحی، ۱۳۸۷). از آن‌جا که روش‌های مبتنی بر نظریهٔ سؤال-پاسخ به حجم نمونهٔ وسیعی برای بررسی سوگیری نیاز دارند، روش‌های پارامتری و غیرپارامتری برای این منظور توسعه یافته‌اند (نارایانان و سوامیناتان، ۱۹۹۶). در نظریه سؤال-پاسخ برای بررسی کارکرد متفاوت سؤال از دو اصطلاح گروه مرجع و گروه هدف استفاده می‌شود. گروه مرجع، گروهی است که بر مبنای مقایسه بوده و پارامترهای به دست آمده برای سایر گروه‌ها که تحت عنوان گروه یا گروه‌های هدف مورد اشاره قرار می‌گیرند با آن مقایسه می‌شوند.

یکی از حوزه‌هایی که مطالعات مربوط به بررسی کارکرد متفاوت سؤال روی آن متمرکز است، بررسی سؤالات و آزمون‌های مربوط به زبان دوم[8] است. منظور از زبان دوم، زبانی است که فرد از طریق آموزش رسمی آن را فرا می‌گیرد. به عنوان مثال افرادی که در امتحان تافل شرکت می‌کنند. شواهد به دست آمده از برخی پژوهش‌ها که برای بررسی کارکرد متفاوت سؤال در آزمون کلمات زبان دوم[9] نشان می‌دهد که اگرچه وجود کارکرد متفاوت برای برخی از سؤالات در یک آزمون، با سوگیری به نفع زنان یا مردان ثابت شده است، ولی شواهد پژوهشی کافی در سوگیری کل آزمون از نظر جنس در دست نیست. با این حال این امکان وجود دارد که وجود تعدادی

1. simulation item bias test (SIBTEST)
2. Shealy & Stout
3. Angof
4. full chi square
5. Scheuneman
6. Mantel and Haenszel
7. Millsap & Everson
8. second language
9. second language vocabulary test

زیادی سؤال دارای سوگیری در یک آزمون نتایج کل آزمون را تحت تاثیر قرار دهد. این موضوع در ساخت بانک سؤال که متشکل از این نوع سؤالات است، اهمیت دارد (تاکالا و کافتانجیوا[1]، ۲۰۰۰). در آزمون‌های زبان خارجی، آزمون‌های بخش گرامر و تلفظ[2] بیشتر از سایر بخش‌های مربوط به آزمون زبان دوم تحت تأثیر ویژگی‌های جمعیت‌شناختی افراد امتحان دهنده قرار می‌گیرند (کیم، ۲۰۰۱).

روش

همان‌طور که قبل از این ذکر شد بررسی کارکرد متفاوت سؤال از جنبه‌های مختلف مورد بررسی قرار گرفته است یکی از رویکردها در این خصوص نظریهٔ سؤال- پاسخ است. براساس نظریهٔ سؤال-پاسخ، کارکرد متفاوت سؤال زمانی روی می‌دهد که احتمال پاسخ درست به یک سؤال برای افراد دارای سطح خصیصهٔ یکسان که متعلق به زیر گروه‌های مختلف جامعه مورد نظر هستند متفاوت است. الگوی یک پارامتری نظریه سؤال-پاسخ جذابیت‌های ریاضی الگوی راش[3] و انعطاف‌پذیری الگوی دوپارامتری را به طور همزمان دارا است. در معادلهٔ این الگو که در زیر ارائه شده احتمال پاسخ درست به سؤال به وسیلهٔ پارامترهای توانایی (θ)، دشواری (β) و مقدار شیب ثابت (α) برای همهٔ سؤال‌ها برآورد می‌شود (همبلتون و سوامینانتان[4]، ۱۹۸۵).

$$P(X=1|\theta,\beta)=\frac{e^{1.7\alpha(\theta-\beta)}}{1+e^{1.7\alpha(\theta-\beta)}}$$

تعریف احتمالی کارکرد متفاوت سؤال در بطن نظریهٔ سؤال-پاسخ به این معنی است که تابع ویژگی سؤال[5] برای زیرگروه‌های جامعه یکسان نیست. از آن‌جا که توابع ویژگی سؤال به طور کامل به وسیلهٔ پارامترهای سؤال و فرد تعیین می‌شوند، پس تحلیل کارکرد متفاوت سؤال را می‌توان هم با مقایسهٔ مستقیم پارامترها و هم مقایسهٔ مساحت بین منحنی‌های ویژگی سؤال در بین گروه‌های مختلف انجام داد. در الگوی یک پارامتری تنها پارامتر دشواری است که تابع ویژگی سؤال را مشخص می‌کند. این مطلب به این معنی است که منحنی ویژگی سؤال در زیرگروه‌ها یکسان خواهد بود، اگر، و فقط اگر، پارامتر دشواری سؤال در آن‌ها یکسان باشد. تحت این شرایط می‌توان از آمارهٔ T برای بررسی معناداری تفاوت دشواری سؤال در بین گروه‌های مختلف استفاده کرد. به طور معمول اگر در یک سؤال ۱/۹۶< /t/ باشد سؤال مورد نظر دارای کارکرد متفاوت

1. Takala & Kaftandjieva
2. pronunciation
3. rasch model
4. Hambleton & Swaminathan
5. item characteristic functions

خواهد بود. عبارات SE_F و SE_R به ترتیب خطای استاندار برآورد برای گروه‌های هدف و مرجع هستند (تاکالا و کافتانجیوا، ۲۰۰۰).

$$T = \frac{b_F - b_R}{\sqrt{SE_F{}^2 + SE_R{}^2}}$$

این روش قادر به تشخیص کارکرد متفاوت یکنواخت است. به این معنی که فرض می‌کند بین سطح توانایی و عضویت در گروه تعامل وجود ندارد و احتمال پاسخ درست به سؤال در همهٔ سطوح خصیصه برای یکی از گروه‌ها بیشتر است. در صورتی‌که بین عضویت در گروه و سطح خصیصه تعامل باشد پارامتر شیب برای گروه‌ها متفاوت خواهد بود و در نتیجه منحنی ویژگی سؤال گروه‌ها یکدیگر را قطع خواهند کرد.

علاوه بر مقایسه دشواری تک‌تک سؤال‌ها در بین گروه‌های مختلف با روش فوق، می‌توان برازش الگو در حالتی که همهٔ داده‌ها به عنوان یک گروه واحد تحلیل می‌شوند و حالتی که داده-های هر یک از گروه‌ها به صورت جداگانه تحلیل می‌شوند را با آزمونLOG LIKELIHOOD 2- نیز مورد بررسی قرار داد (کیم، ۲۰۰۱). در این روش از آمارهٔ LOG LIKELIHOOD 2- برای مقایسهٔ بین برازش الگوی شناسایی کارکرد متفاوت سؤال و الگوی بدون شناسایی کارکرد متفاوت سؤال استفاده می‌شود. سپس تفاوت آن‌ها که دارای توزیع کای‌دو با درجه آزادی برابر تفاوت بین تعداد پارامترها در الگوی بدون شناسایی کارکرد متفاوت سؤال و الگوی شناسایی کارکرد متفاوت سؤال است (که در نهایت برابر با تعداد سؤالات خواهد بود) برای ارزیابی برازش الگوی استفاده می‌شود. اگر کای‌دو حاصل از کای‌دو بحرانی با همین درجه آزادی بزرگتر باشد الگو دارای شناسایی کارکرد متفاوت سؤال از الگوی بدون شناسایی کارکرد متفاوت سؤال برازش بهتری دارد.

روش‌های توصیف‌شده در فوق با استفاده از نرم‌افزار بای‌لوگ-امجی که در پژوهش فعلی برای تحلیل از آن استفاده شده، قابل اجرا است. در این پژوهش برای برآورد پارامترها از تابع اجایو نرمال[1] استفاده شده و متغیر گروه‌بندی برای تحلیل پاسخ‌های افراد، جنس است. فایل فرمانی که داده‌های تحلیل حاضر به وسیله آن برای بررسی کارکرد متفاوت سؤال تحلیل شدند در آخر مقاله ارائه شده است که علاقه‌مندان می‌توانند با بررسی و مطالعهٔ آن داده‌های خود را تحلیل کنند. برای آشنایی بیشتر با نرم افزار بای‌لوگ-امجی به پیوست کتاب *نظریه‌های جدید روان-سنجی برای روان‌شناسان* تألیف امبرستون و رایز (۲۰۰۰)و روپ[2] (۲۰۰۳) مراجعه کنید.

1. normal ojive function
2. Rupp

در تحلیل حاضر، داده‌های خرده‌مقیاس گرامر مربوط به دفترچهٔ آزمون زبان تخصصی انگلیسی در آزمون سراسری سال ۱۳۸۴ استفاده شد. کل داده‌ها ۲۲۲۹۴ هزار نفر بودند. چون داده‌ها حاوی اطلاعات مربوط به پاسخ‌های افراد برای سایر زبان‌ها نیز بود، پس این اطلاعات از کل داده‌ها حذف شدند و فقط افرادی که زبان امتحانی آن‌ها انگلیسی بود انتخاب شدند. بنابراین، تعداد کل آزمودنی‌ها از ۲۲۲۹۴ به ۲۲۱۹۴ مورد کاهش یافت. سپس با انتخاب آزمودنی‌های دارای دیپلم ریاضی، انسانی و تجربی حجم داده‌ها به ۲۱۹۳۹ مورد رسید. در نهایت نیز با انتخاب آزمودنی‌هایی که حوزهٔ امتحانی آن‌ها در داخل کشور بود، حجم نهایی به ۲۱۹۲۴ تقلیل یافت که از این تعداد ۷۷۴۳ نفر مرد و ۱۴۱۸۱ نفر زن بودند. با توجه به هدف پژوهش، یعنی بررسی کارکرد متفاوت سؤال‌ها براساس جنس آزمودنی‌ها، براساس فراوانی کل جامعه در این متغیر، از هر یک از این گروه‌ها یک نمونهٔ تصادفی ۳۰۰۰ انتخاب شد.

یافته‌ها

ابتدا پارامترها برای هر یک از گروه‌ها به صورت جداگانه برآورد می‌شوند. چون الگوی یک پارامتری است، فقط پارامتر دشواری سؤالات تغییر خواهد کرد. این پارامترها برای هر یک از دو گروه مرجع (زنان) و کانونی (مردان) در جدول ۲ ارائه شده‌اند. به علاوه پارامتر شیب همهٔ سؤالات برابر با ۰/۴۶۹ برآورد شده است. متوسط برآورد دشواری سؤال‌ها و انحراف استاندارد آن به ترتیب در گروه زنان ۰/۰۴۶ و ۱/۳۷۷ و در گروه مردان ۰/۰۳۷ و ۱/۳۷۱ است. میانگین دشواری تعدیل شده برای گروه کانونی برابر با ۰/۰۰۸- برآورد شده است. این میانگین با کم کردن میانگین دشواری گروه کانونی از میانگین دشواری گروه مرجع به دست می‌آید، در حالی که میانگین گروه مرجع برابر صفر در نظر گرفته می‌شود (دوتویت[1]، ۲۰۰۳). در واقع با این کار توزیع پارامترهای دشواری بر روی مقیاس یکسانی قرار می‌گیرد. سپس این میانگین تعدیل شده، از تمام مقادیر دشواری گروه کانونی کم شده تا پارامترهای تعدیل شده برای این گروه به دست آیند. سپس مقادیر دشواری تعدیل شده گروه کانونی از مقادیر دشواری گروه مرجع کم شده و به این ترتیب تفاوت‌های بین‌گروهی در دشواری سؤال‌ها به دست می‌آید. این محاسبات در جدول۲ ارائه شده‌اند. همان‌طور که دیده می‌شود تنها تفاوت گروه‌ها در سؤال ۸ از t بحرانی (۱/۹۶) بزرگتر است. در بقیهٔ موارد بین مقدار دشواری گروه‌ها از نظر آماری تفاوت معناداری وجود ندارد.

1. DuToit

جدول۲. پارامترهای دشواری برآورد شده در گروه مرد و زن و معناداری آن‌ها

سؤال	زن	مرد	دشواری‌های تعدیل شده در گروه مردان	تفاوت	t مشاهده شده
۱	(۰/۰۵۳) -۰/۲۷	(۰۵۱) -۰/۱۶	-۰/۱۵۲ (۰/۰۵۱)	۰/۱۱۸ (۰/۰۷۴)	۱/۵۹۷
۲	(۰/۰۵۳) ۰/۲۱۶	(۰/۰۵) ۰/۰۷۶	۰/۰۸۴ (۰/۰۵)	-/۱۳۲ (۰/۰۷۳)	-۱/۷۹
۳	(۰/۰۷) ۲/۳۷۸	(۰/۰۶۹) ۲/۲۹۴	۲/۳۰۲ (۰/۰۶۹)	-/۰۷۷ (۰/۰۹۹)	-۰/۷۸
۴	(۰/۰۵۵) ۰/۸۴۱	(۰/۰۵۳) ۰/۹۴۸	۰/۹۵۶ (۰/۰۵۳)	۰/۱۱۵ (۰/۰۷۶)	۱/۵۲
۵	(۰/۰۵۳) ۰/۴۷	(۰/۰۵۲) ۰/۴۳۸	۰/۴۴۶ (۰/۰۵۲)	-/۰۲۴ (۰/۰۷۴)	-۰/۳۲
۶	(۰/۰۶۹) -۲/۳۷۴	(۰/۰۵) -۲/۲۶۸	-۲/۲۶ (۰/۰۵)	۰/۱۱۴ (۰/۰۸۵)	۱/۳۲
۷	(۰/۰۵۵) ۰/۸۶۶	(۰/۰۵۲) ۰/۹۰۸	۰/۹۱۶ (۰/۰۵۲)	۰/۰۵ (۰/۰۷۶)	۰/۶۵
۸	(۰/۰۵۳) -/۴۱۳	(۰/۰۵۱) -/۵۶۸	-/۵۶ (۰/۰۵۱)	-/۱۴۷ (۰/۰۷۴)	-۲/۰۰*
۹	(۰/۰۵۴) ۰/۶۰۲	(۰/۰۵۲) ۰/۶۶۴	۰/۶۷۲ (۰/۰۵۲)	۰/۰۷ (۰/۰۷۵)	۰/۹۴
۱۰	(۰/۰۶۲) -۱/۸۶۲	(۰/۰۵۱) -۱/۹۵۶	-۱/۹۴۸ (۰/۱۰۵)	-/۰۸۶ (۰/۰۸)	-۱/۰۸۲

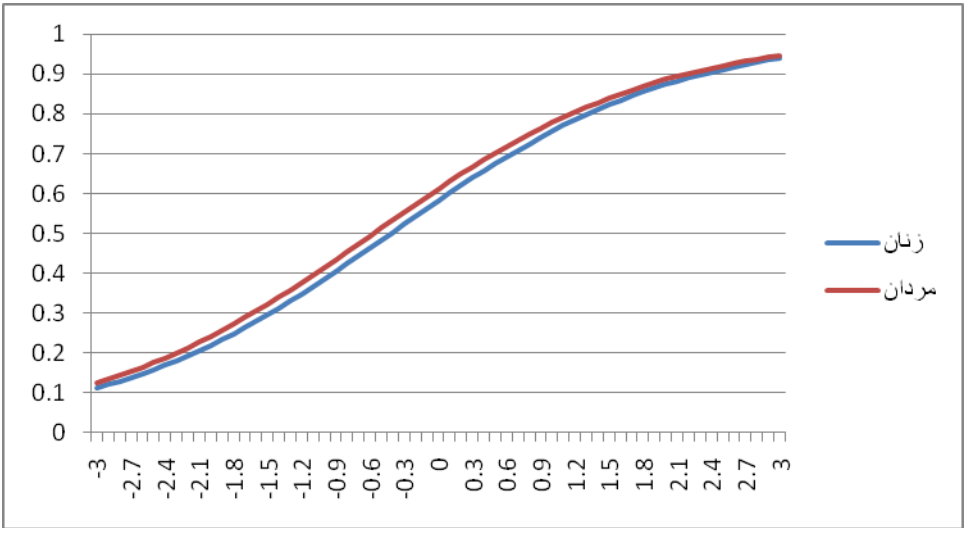
*مقادیر داخل پرانتز خطای استاندارد هر پارامتر است.

برای مقایسه برازش الگوها، مقدار به‌دست آمده برای LOG LIKELIHOOD -2 در آخرین چرخش هر دو الگو را از هم کم می‌کنیم. تفاوت آن‌ها که دارای توزیع کای‌دو با درجهٔ آزادی برابر با تعداد سؤالات است عبارت است از:

$\chi^2 =$ ۷۰۹۶۰/۲۴۲۵- ۷۰۹۳۲/۳۸۹۴ =۲۳/۸۵

از آن‌جا که این مقدار از کای‌دو جدول در سطح ۰/۰۵ (۱۸/۳۱) بزرگتر است می‌توان برازش بهتر الگوی شناسایی کارکرد متفاوت سؤال پذیرفت. در نتیجه می‌توان گفت بین علمکرد پسران و دختران در این سؤال‌ها تفاوت وجود دارد و به عبارتی نمودار ویژگی سؤال این دو گروه حداقل در سؤال ۸ که مقدار t برای آن معنادار شد، تفاوت معناداری با یکدیگر دارد. در واقع سؤال ۸ برای مردان آسان‌تر از گروه زنان بوده است. لازم به ذکر است که معمولاً در روش‌هایی که در این‌جا توصیف شدند ابتدا با روش تفاوت معناداری بین دشواری‌ها در دو گروه مقایسه‌ها صورت

می‌گیرد و سپس با آزمون کای‌دو برازش الگوی‌ها نیز بررسی می‌شود. که اگر این آزمون برازش بهتر بودن الگوی شناسایی کارکرد متفاوت سؤال را نشان داد پس وجود کارکرد متفاوت سؤال‌ها یا به عبارتی وجود سوگیری در آن‌ها پذیرفته می‌شود. در غیر این صورت می‌توان گفت وجود هرگونه شناسایی کارکرد متفاوت سؤال در داده‌ها به لحاظ آماری معنادار نیست (امبرستون و رایس، ۲۰۰۰؛ دوتویت، ۲۰۰۳). بررسی منحنی ویژگی سؤال ۸ دو گروه که در نمودار ۱ ارائه شده، نشان می‌دهد تقریباً در تمامی سطوح خصیصه، به خصوص در نواحی میانی خصیصه، احتمال پاسخ درست برای مردان در مقایسه با زنان بیشتر بوده است.



نمودار۱. منحنی ویژگی سؤال ۸ برای هر دو گروه مرد و زن

بحث و نتیجه‌گیری

همان‌طور که گفته شد، هدف اصلی این مقاله بررسی کارکرد متفاوت سؤال با استفاده از الگوی یک پارامتری در نرم افزار بای‌لوگ-ام‌جی بود. در حال حاضر، این نرم‌افزار فقط قادر به بررسی کارکرد متفاوت سؤال با الگوی یک پارامتری است. به علاوه چون الگو، یک پارامتری است این روش فقط قادر به بررسی کارکرد متفاوت سؤال از نوع یکنواخت است. مزیت این الگو، استحکام الگوی نظری و تأکید آن بر پارامتر دشواری است و نقطه ضعف این الگو آن است که بسیاری از سؤال‌ها با الگو برازش ندارند و همین موضوع می‌تواند تأثیر منفی بر برآورد دشواری داشته باشد. به هر حال سؤالاتی که با این روش دارای کارکرد متفاوت باشند، باید با سایر روش‌های معتبر نیز تحلیل شده و سپس از نظر محتوایی تجزیه و تحلیل شوند و مورد بازنگری قرار گیرند. در تحلیل

حاضر، سؤال ۸ یکی از این سؤالاتی است که لازم است علاوه بر تحلیل با سایر روش‌ها به لحاظ محتوایی تحلیل شده و مورد بازنگری قرار گیرد. محتوای سؤال ۸ عبارت است از:

Somebody ……. Jack phoned while you were out.

1) named 2) naming 3) being named 4) which named

پاسخ گزینهٔ یک است و معنای کلی جمله عبارت است از: *کسی به نام جک، موقعی که تو نبودی، زنگ زد.*

در چارچوب نظریهٔ کلاسیک و نقطه‌ضعف‌های موجود در آن بررسی کارکرد متفاوت سؤال، امری نادرست است، زیرا ویژگی متغیربودن آماره‌های سؤال و افراد از یک گروه به گروه دیگر به طور ذاتی در این نظریه وجود دارد (دولیس[1]، ۲۰۰۶). به احتمال زیاد علت اصلی کارکرد متفاوت سؤال، که به معنی فقدان نامتغیر بودن پارامترها است، چندبعدی بودن زیر مجموعه‌هایی از سؤال‌ها است که از نظر محتوایی بسیار مشابه‌اند. به این معنی که هر چند همهٔ سؤال‌های موجود در یک مقیاس می‌توانند نشانه‌های خوبی از یک عامل مشترک باشند، ممکن است برخی از سؤال‌ها به وسیله ابعاد مزاحم[2] که اهمیت کمتری دارند تحت تأثیر قرار بگیرند. در نتیجه تفاوت میانگین گروه‌های مختلف در این عامل‌های مزاحم، هنگامی که سؤال‌ها تحت چارچوب تک‌بعدی مطالعه می‌شوند، ممکن است به شناسایی کارکرد متفاوت سؤال منجر شود. این نوع شرایط در سنجش روان‌شناختی بیشتر رایج است، که در آن پیدا کردن یک مقیاس مورد علاقه که به طور واقعی فقط و فقط یک صفت مکنون را اندازه‌گیری کند واقعاً نادر است. این واقعیت اندازه‌گیری روان‌شناختی، در کل استفاده از الگوی‌های تک‌بعدی نظریهٔ سؤال-پاسخ و به طور خاص روش-های تعیین کارکرد متفاوت سؤال مبتنی بر نظریهٔ سؤال-پاسخ را مورد سؤال قرار می‌دهد (ناناداکومار[3]، ۱۹۹۱). بنا به گفته امبرستون و رایس (۲۰۰۰) روش‌های جدید تحلیل کارکرد متفاوت سؤال، نظیر الگوی چندبعدی شلی- استوت[4] (یک روش غیرپارامتریک) این امکان را فراهم می‌کند تا ابعاد مزاحم کوچکتر با تحلیل کارکرد متفاوت سؤال یکپارچه شود. روس و استوت[5] چگونگی تلفیق این روش با نظریهٔ بنیادی و مزایای حاصل از آن را توضیح داده‌اند و برنامه‌های کامپیوتری که سیب‌تست و مالتی‌سیب[6] نامیده می‌شوند، برای اجرای این طبقه از روش‌ها در دسترس‌اند. نتایج پژوهش‌هایی که تاکنون با این روش‌ها انجام گرفته است بسیار نوید بخش‌اند. مهمتر این‌که، نظریهٔ زیربنایی این الگو با آزمون‌های دنیای واقعی همسو است و انتظار

1. Devellis
2. nuisance
3. Nandakumar
4. Shealy – Stout multidimensional model
5. Roussos & Stout
6. SIBTEST & MULTISIB

می‌رود که در سال‌های آینده این چارچوب، نظر پژوهشگران را بیشتر به خود جلب کند. به عنوان آخرین نکته، بررسی کارکرد متفاوت سؤال، تنها محدود به حوزه آزمون‌های پیشرفت تحصیلی نیست بلکه سایر ابزارها و مقیاس‌های سنجش روانی را نیز شامل می‌شود.

منابع

امبرستون، سوسان.، و رایز، استیون. (۲۰۰۰). *نظریه‌های جدید روان‌سنجی برای روان‌شناسان*، ترجمهٔ حسن پاشا شریفی، ولی‌الله فرزاد، مجتبی حبیبی و بلال ایزانلو، ۱۳۸۸. تهران: انتشارات رشد.

فتوحی، لیلا. (۱۳۸۷). *بررسی کارکرد افتراقی سؤال در سؤالات کنکور کارشناسی ارشد رشته روان‌شناسی سال ۸۴*. پایان‌نامهٔ کارشناسی ارشد رشتهٔ سنجش و اندازه‌گیری، دانشگاه علامه طباطبایی، دانشکدهٔ روان-شناسی و علوم تربیتی.

Berk, R, A. (1982). *Handbook of Methods for Detecting Test Bias*, Baltimore: Johns Hopkins University Press.

Devellis, R. F. (2006). Classical test theory. *Medical care*, 44: 50-59.

Dutoit, M. (2003). IRT FROM SSI: BILOG-MG, MULTILG, PARSCALE, TESTFACT. Scientific Software international, Inc.

Embereston, S. E., & Reise, S. P. (2000). *Item Response Theory for psychologists*. London: Lawrence Erlbaum Associates.

Hambleton, R.K., & Swaminathan, H. (1985). *Item Response Theory: principles and applications.* Boston, MA: Kluwer-Nijhoff.

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing*, 18: 89-114.

Millsap, R., & Everson, H. T. (1993). Methodology Review: statistical Approaches for assessing Measurement bias. *Applied psychological Measurement.* 17: 297–334.

Nandakumar, R. (1991). Traditional Dimensionality Versus Essential Dimensionality, Journal of Educational Measurement. 28(2): 99-11 7.

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

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): 517–529.

Rupp, A. A. (2003). Item Response Modeling With BILOG-MG and MULTILOG for Windows, *International Journal of Testing.* 3(4): 365–384.

Sireci, S, G., & Allalouf, A. (2003) Appraising item equivalence across multiple languages and cultures, *Language Testing* .20:148-166.

Takala, S., & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing*, 17: 323-340.

فایل فرمان استفاده شده برای تحلیل حاضر

```
>COMMENT
DIF ANALYSIS FOR GRAMMER SECTION
>GLOBAL DFName = 'merging.dat',
       NPArm = 1,
       SAVe;
>SAVE CALib = 'dif.CAL',
     PARm = 'dif.PAR',
     SCOre = 'dif.SCO',
     TSTat = 'dif.TST',
     ISTat = 'dif.IST',
     DIF = 'dif.DIF';
>LENGTH NITems = (10);
>INPUT NTOtal = 10,
      NIDchar = 6,
      NGRoup = 2,
      KFName = 'key.DAT',
      DIF;
>ITEMS ;
>TEST1 TNAme = 'TEST0001',
      INUmber = (1(1)10);
>GROUP1 GNAme = 'GROUP001',
       LENgth = 10,
       INUmbers = (1(1)10);
>GROUP2 GNAme = 'GROUP002',
       LENgth = 10,
       INUmbers = (1(1)10);
(6A1, I1, 10A1)
>CALIB CYCles = 200,
      NEWton = 100,
      CRIt = 0.0050,
      PLOt = 1.0000,
      NOSprior,
      NOFloat,
      CHIsquare = (10, 13);
>SCORE ;
```