فصلنامه اندازه‌گیری تربیتی
دانشگاه علامه طباطبائی
سال هشتم، شماره ۳۲، تابستان ۹۷، ص ۴۱ تا ۶۲

Quarterly of Educational Measurement
Allameh Tabataba'i University
Vol. 8, No. 32, Summer 2018

کنش افتراقی جنسیتی سوالات آزمون کنکور سراسری کارشناسی ارشد رشته‌ی آموزش زبان انگلیسی در ایران: مقایسه‌ی روش‌های آماری رگرسیون لجیستیک و منتل-هنسزل

معصومه استاجی[۱]، نگار بابانژاد کفشگر[۲]

تاریخ دریافت: ۹۷/۰۱/۲۳
تاریخ پذیرش: ۹۷/۰۳/۰۵

چکیده

این مطالعه قصد داشت تا در آزمون بزرگی چون کنکور کارشناسی ارشد آموزش زبان انگلیسی در ایران، سوالاتی که شامل کنش افتراقی جنسیتی باشند را با استفاده از دو نمونه از کاربردی ترین و مفید ترین روش‌ها (رگرسیون لجیستیک و منتل-هنسزل) پیدا کند. به علاوه، یک تحلیل محتوایی از سوالات پیدا شده‌ی شامل کنش افتراقی انجام شد تا منابع زبانی چنین سوگیری‌هایی مشخص شوند. داده‌ها از پایگاه داده‌های سازمان سنجش ایران به دست آمد. پاسخنامه‌های ۲۲۱۷ زن و ۷۳۵ مرد در کنکور سال ۲۰۱۵ برای یافتن سوالاتی که شامل رفتارهای انحرافی بودند بررسی شدند. نتایج حاصل از روش رگرسیون لجستیک نشان داد که هشت سوال شامل کنش افتراقی بودند. چهار سوال از این سوالات به نفع آقایان و چهار سوال هم به نفع خانم‌ها سوگیر بودند. یازده سوال توسط روش منتل-هنسزل به عنوان سوالات شامل کنش افتراقی شناخته شدند که شش مورد از آن‌ها به نفع آقایان و پنج مورد به نفع خانم‌ها سوگیری داشتند. تحلیل محتوایی سوالات شامل کنش افتراقی، هیچ دلیل زبانی خاصی را برای رفتار افتراقی سوالات نشان نداد.

واژگان کلیدی: سوگیری جنسیتی، کنش افتراقی سوال، رگرسیون لجیستیک، منتل-هنسزل، کنکور کارشناسی ارشد رشته آموزش زبان انگلیسی، عدالت آزمون

---

۱. دانشیار آموزش زبان انگلیسی دانشکده ادبیات فارسی و زبان‌های خارجی، دانشگاه علامه طباطبائی، تهران، ایران. (نویسنده مسئول) mestaji74@gmail.com

۲. کارشناسی ارشد آموزش زبان انگلیسی، دانشکده ادبیات فارسی و زبان‌های خارجی، دانشگاه علامه طباطبائی، تهران، ایران.

مقدمه

عدالت آزمون که در حوزه‌ی آزمون‌سازی به عنوان یک مسئله‌ی مهم روان‌سنجی و آموزشی شناخته می‌شود (کمیلی[1]، ۲۰۰۶)، توسط استانداردهای آزمون‌های آموزشی و روانشناسی بدین گونه تعریف شده است: عدم وجود هرگونه سوگیری در فرآیندهای آزمون مانند طراحی آزمون، توسعه‌ی آزمون، برگزاری آزمون، و مراحل نمره‌دهی (ای‌ای‌آرای، ای‌پی‌ای، و ان‌سی‌ام‌ای[2]، ۱۹۹۹). آزمونی که از نظر ویژگی‌های آماری و روانسنجی نامربوط به هدف آزمون نسبت به یکی از دو گروه شرکت‌کنندگان سوگیری نشان دهد به عنوان یک آزمون سوگیر شناخته می‌شود (آنگوف[3]، ۱۹۹۳).

در بسیاری از کشورها مانند ایران، کره، و ژاپن، راهیابی به دانشگاه‌های سراسری منوط به قبولی در آزمون ورودی دانشگاه است و با توجه به اینکه ظرفیت دانشگاه‌های سراسری محدود می‌باشد منطقی است که افراد شایسته تر به این دانشگاه‌ها راه پیدا کنند. در سیاست آموزشی این چنینی نقش یافتن و حذف سوگیری آزمون بسیار مهم و قابل توجه است. به طور عمده، سوالات آزمون درصورتی سوگیر می‌شوند که منابع دشواری آن‌ها با توانایی مشخص مورد سنجش آزمون ارتباطی نداشته باشند و این عوامل نامربوط بر عملکرد شرکت‌کنندگان آزمون تاثیر بگذارند (آنگوف، ۱۹۹۳). این عوامل نامربوط شامل عواملی نظیر جنسیت، پیشینه‌ی آموزشی، سن، و ملیت هستند. نوربخش (۱۳۸۹) و خدایی (۱۳۸۸) نیز عواملی چون پیشینه‌ی فرهنگی، اجتماعی و اقتصادی خانوادگی را موثر در تفاوت موفقیت تحصیلی افراد می‌دانند. بنابراین، در صورتی آزمون سوگیر است که سوالات آن بیشتر از هدف اصلی خود، در حال سنجش این عوامل نامربوط باشند. اگر دو گروه متفاوت از شرکت‌کنندگان آزمون به صورت منظم و مشخصی در عملکردشان تفاوت نشان دهند، منابع این تفاوت باید به صورت کامل مورد بررسی و تحلیل قرار گیرند. چرا که این منابع تهدیدی بالقوه برای اعتبار تفاسیر نمرات آزمون به شمار می‌روند (بیرجندی و امینی، ۲۰۰۷).

لازم به ذکر است که عدالت آزمون می‌بایست برای شرکت‌کنندگان برقرار شود تا تفاسیر نمرات آزمون معتبر باشند. کنش افتراقی (DIF) زمانی اتفاق می‌افتد که احتمال پاسخگویی صحیح به سوال در دو گروه شرکت‌کنندگان با توانایی برابر، متفاوت باشد

1. Camili
2. AERA, APA, & NCME
3. Angoff

(کمیلی و شپارد[1]، ۱۹۹۴). وجود کنش افتراقی به این معنی است که عوامل مربوط به عضویت در گروهی خاص، احتمال پاسخ صحیح را تحت تاثیر قرار می‌دهند و بنابراین ارزیابی عادلانه را تهدید می‌کنند. به همین علت، غربالگری DIF بخشی ضروری در توسعه‌ی آزمون است، و با توجه به تاثیر منفی سوالات دارای DIF بر افراد شرکت‌کننده، اهمیت این غربالگری نمی‌تواند کوچک شمرده شود (پائه[2]، ۲۰۰۴). در حقیقت وجود DIF یک شرط لازم برای سوگیری است ولی کافی نیست (مک‌نامارا و رووور[3]، ۲۰۰۶). از آنجائی که روش‌های یافتنDIF براساس مقایسه‌ی گروه‌هایی با توانایی برابر اجرا می‌شوند، مقدار بزرگی از DIF نشان می‌دهد که سوال مورد نظر درحال سنجش موارد اضافی است که کارکردشان در یک گروه نسبت به گروهی دیگر متفاوت است (آنگوف، ۱۹۹۳؛ کمیلی و شپارد، ۱۹۹۴؛ روسوس و استات[4]، ۱۹۹۶). در حال حاضر چهار روش برای تشخیص DIF استفاده می‌شود: مانتل هانسزل، سیبتست، رگرسیون لجستیک و نظریه پاسخ سوال (IRT).

یکی از معروفترین روش‌های تشخیص DIF منتل- هنسزل است که به طور گسترده‌ای در مطالعات DIF مورد استفاده قرار گرفته است (هلند و تایر[5]، ۱۹۸۸). در این روش DIF با استفاده از جدول سه‌گانه توافقی تشخیص داده می‌شود. یکی دیگر از روش‌های اندازه‌گیری DIF، رگرسیون لجستیک است که تاثیر گروه در سوالات دو جانبه (که به صورت پاسخ صحیح یا غلط نمره داده می‌شوند) را مورد بررسی قرار می‌دهد (سوامیناتان و راجرز[6]، ۱۹۹۰). همچنین برای سوالات چندجانبه هم می‌توان از این روش استفاده کرد (فرنچ و میلر[7]، ۱۹۹۶). دلیل دیگر برای محبوبیت رگرسیون لجیستیک این است که اجازه می‌دهد مدل‌سازی DIF هماهنگ و ناهماهنگ نسبت به روش IRT کمتر نیاز به محاسبات پیچیده داشته باشد. بررسی عدالت آزمون در کنکور سراسری کارشناسی ارشد رشته آموزش زبان انگلیسی برای شرکت‌کنندگان بسیار مهم است چون تاثیر بسزایی بر آینده‌ی آن‌ها می‌گذارد. همه‌ی شرکت‌کنندگان می‌خواهند مطمئن باشند این توانایی آن‌ها است که مورد ارزیابی

1. Camili & Shepard
2. Pae
3. Mcnamara & Roever
4. Roussos & Stout
5. Holland & Thayer
6. Swami Nathan & Rogers
7. French & Miller

قرار می‌گیرد و هیچ عامل دیگری در کار نیست. مطالعه‌ی حاضر با استفاده از روش های منتل-هنسزل و رگرسیون لجیستیک قصد داشت تا در قسمت زبان عمومی کنکور سراسری کارشناسی ارشد رشته‌ی آموزش زبان انگلیسی سوالاتی را که شامل کنش افتراقی نسبت به گروه جنسی خاصی (زن یا مرد) می‌باشند پیدا کند. به طور خاص، بخش عمومی‌زبان انگلیسی کنکور سراسری کارشناسی ارشد رشته آموزش زبان انگلیسی ایران برای DIF جنسیت مورد بررسی قرار گرفت تا در مورد عدالت آزمون اطمینان حاصل شود و منابع سوگیری آزمون (در صورت وجود) شناسایی شوند.

مطالعات متفاوتی در حوزه ی آزمون‌سازی زبان وجود دارد که هدفشان ایجاد شرایط عادلانه در آزمون‌ها و جلوگیری از سوگیری آن ها بوده است. در مطالعات آزمون‌سازی زبان، تحقیقات زیادی وجود DIF را در میان گروه‌های مختلف جنسی مورد بررسی قرار دادند. برای مثال، رایان و بکمن[1] (۱۹۹۲) مطالعه‌ای را برای یافتن کنش افتراقی جنسیتی (gender DIF) در آزمون‌های تافل (TOEFL) و اف سی ای(FCE) ترتیب دادند. آن‌ها متوجه شدند که مردان و زنان تفاوت معنی‌دار خاصی در عملکردشان در پاسخگویی به سوالات از خود نشان ندادند. وینر و لوخل[2] (۱۹۹۷) هم اظهار داشتند که بخش‌های درک مطلب آزمون تافل اساساً از نظر جنسیتی کنش متفاوت خاصی را نشان نمی‌دهند. همچنین لین و وو[3] (۲۰۰۳) در تحقیق خود در مورد آزمون EPT در چین، به نتایج مشابهی رسیدند و هیچ کنش افتراقی جنسیتی را تشخیص ندادند. به هر حال، نتایج متفاوتی هم توسط محققان دیگر به دست آمد. برای مثال کارلتون و هریس[4] (۱۹۹۲) از روش های منتل-هنسزل و رگرسیون لجیستیک برای یافتن DIF استفاده کردند و نتایج تحقیقشان نشان داد که زنان نسبت به مردان در پاسخگویی به سوالات مربوط به زیبایی‌شناسی و روابط انسانی موفق‌تر عمل کردند. ولی در پاسخگویی به سوالات علمی‌تر و موارد اجرایی، زنان عملکرد ضعیفتری نسبت به مردان داشتند. آزمون SAT هم از لحاظ کنش افتراقی جنسیتی توسط محققان دیگری مانند لارنس، کرلی و مک هیل[5] (۱۹۹۸) و لارنس و کرلی (۱۹۸۹) مورد بررسی واقع شد. آن‌ها از روش استانداردسازی برای یافتن DIF استفاده کردند و متوجه شدند که

1. Ryan & Bachman
2. Wainer & Luckele
3. Lin & Wu
4. Carlton & Harris
5. Lawrence, Curly, & McHale

در سوالات مربوط به خواندن متون فنی، زنان به صورت قابل توجهی کمتر از همتایان مردشان موفق هستند.

کنولی[1] (۲۰۰۳) تحقیقی را انجام داد که کارشناسان را در مورد قضاوت‌های نا به جا در مورد وجود سوالات DIF و پیش‌بینی آن‌ها هشدار می‌دهد (همان‌طور که در رضایی و شعبانی، ۱۳۸۸ نقل شده است). او به ما پیشنهاد می‌کند که روی نظرات شخصی کارشناسان در مورد مطالعات یافته‌های سوگیری خیلی حساب نکنیم. همان‌طور که رضایی و شعبانی (۱۳۸۸) گزارش کردند مهم ترین نکته در پیشنهاد کنولی (۲۰۰۳) این است که اگر مردان و زنان به صورت متفاوتی به سوال پاسخ دادند، این به این معنی نیست که DIF وجود دارد و یا اینکه سوگیری‌اش مشخص است.

در ایران مطالعاتی وجود دارد که عواملی چون جنسیت، رشته‌ی دانشگاهی، سن و غیره را به عنوان متغیر اصلی شان قرار داده‌اند. این مطالعات توسط محققانی انجام شده است که آزمون‌هایی مثل آزمون مهارت زبان دانشگاه تهران (UTEPT) را بررسی نموده‌اند (علوی و کرمی، ۲۰۱۰؛ کرمی، ۲۰۱۱؛ امیریان، ۲۰۱۲؛ رضایی و شعبانی، ۱۳۸۸).

برای مثال امیریان (۲۰۱۲) از روش‌های منتل-هنسزل و رگرسیون لجیستیک برای یافتن DIF جنسیتی در میان سوالات آزمون UTEPT استفاده کرد. شرکت‌کنندگان در این مطالعه ۸۹۹ مرد و ۶۵۱ زن بودند که به ترتیب به عنوان اعضای گروه مرجع و کانونی در نظر گرفته شدند. او برای اجرا کردن روش منتل-هنسزل از برنامه‌ی دیفاز (DIFAS) استفاده نمود و برای اجرا کردن رگرسیون لجیستیک، برنامه‌ی اس‌پی‌اس‌اس ناگلکرک (Nagelkerke's SPSS Syntax) را به کار برد. از میان ۱۰۰ سوال، طبق نتایج منتل-هنسزل، ۳۱ سوال DIF را نشان دادند. اگرچه، هیچ کدامشان نوع C (مقدار بزرگ) از DIF را نشان ندادند. همه‌ی آن سوالات یا دارای مقدار بی‌اهمیت یا متوسط DIF بودند. نتایج رگرسیون لجیستیک هم نشان دادند که ۲۹ سوال شامل DIF بودند که ۲۵ عدد از آن‌ها هماهنگ و ۴ عدد از آن‌ها ناهماهنگ بودند.

مفاهیم DIF هماهنگ و ناهماهنگ توسط ملنبرگ[2] (۱۹۸۲) تعریف شد: DIF هماهنگ در صورت عدم وجود تعامل بین عضویت گروه و سطح توانایی وجود دارد؛ و

1. Conoley
2. Mellenbergh

درصورتی سوال DIF ناهماهنگ است که رابطه‌ای بین عضویت در گروه و سطح توانایی وجود داشته باشد (همان‌طور که در سوامیناتان و راجرز،۱۹۹۰ آمده است) در واقع تمامی-سوالات در سطح A بودند یعنی دارای مقدار جزئی از DIF بودند که این موضوع نشان می‌دهد این آزمون سوگیر نبوده است.

در مطالعه‌ی دیگری در مورد آزمون UTEPT، کرمی(۲۰۱۱) گزارش داد که از ۱۹ سوال شامل DIF که توسط روش IRT شناسایی شده بودند ۷ سوال به نفع مردان سوگیری داشتند، در حالیکه ۱۲ سوال به نفع شرکت‌کنندگان زن سوگیری داشتند.

با توجه به تحلیل محتوایی سوالات شامل DIF، باید توجه داشت که در مطالعات بین‌المللی توجه بسیار کمی‌به منابع احتمالی DIF شده است و مطالعات اندکی به کشف این منابع پرداختند (به عنوان مثال اللوف، همبلتون و سیرسی[1]، ۱۹۹۹). اشمیت، هلند و دورانز[2] (۱۹۹۲) سه دلیل برای این واقعیت بیان کردند. اولاً، آن‌ها متوجه شدند که مطالعات DIF همچنان نوپا هستند. تا این لحظه، تمرکز اصلی تحقیقات بر تکنیک‌ها و روش‌های آماری تشخیص DIF بوده است. دوم اینکه، تشخیص DIF برای گروه‌های مختلف منجر به نظریه‌های پیش‌داورانه در مورد سوالات شامل DIF می‌شود. سوماً، بررسی منابع DIF، کار بسیار پیچیده‌ای است، چرا که احتمال بودن همزمان بیش از یک دلیل برای یک سوال شامل DIF وجود دارد. در اینجا، به برخی از مطالعاتی که محتوای سوالات شامل DIF را تجزیه و تحلیل کرده‌اند اشاره شده است. به عنوان مثال کونان و وینستین-شر[3](۱۹۹۰) با استفاده از روش های IRT مطالعه‌ای را در بین دو گروه زبان مادری و جنسیتی در یک آزمون تعیین سطح زبان انگلیسی به عنوان زبان خارجه انجام دادند و همچنین سوالات شامل DIF را از نظر علت‌های زبانی پشت سوگیری آن‌ها بررسی کردند. از مجموع ۳۶ سوال شامل DIF شناسایی شده، توانستند برای علت‌های زبانی ۲۲ سوال فرضیه بسازند در حالی که هیچ نظریه‌ای برای ۱۴ سوال باقیمانده وجود نداشت. در یک مطالعه‌ی دیگر، امیریان (۲۰۱۲) هیچ منبع زبانی خاصی را برای سوالات شناسایی شده‌ی شامل DIF در آزمون UTEPT یافت نکرد.

1. Allalouf, Hambleton, & Sireci
2. Schmitt, Holland, & Dorans
3. Kunnan, & Weinstein-SHR

این مطالعه قصد داشت تا با استفاده از دو روش آماری منتل-هنسزل و رگرسیون لجیستیک سوالات بخش زبان عمومی‌انگلیسی آزمون سراسری کارشناسی ارشد رشته ی آموزش زبان انگلیسی در سال ۲۰۱۵ را از لحاظ عدالت آزمون مورد بررسی قرار دهد. با توجه به اهداف پژوهش، سوالات تحقیق این مطالعه به شرح ذیل مطرح شدند:

۱. آیا سوالات بخش زبان عمومی‌انگلیسی آزمون سراسری کارشناسی ارشد رشته‌ی آموزش زبان انگلیسی نشان دهنده ی مقدار قابل توجهی از DIF (کنش افتراقی سوال) به نفع گروه جنسی خاصی از شرکت‌کنندگان (مردان یا زنان) هستند؟

۲. آیا نتایج روش‌های آماری به کارگرفته شده برای یافتن DIF که همان روش‌های منتل هنسزل و رگرسیون لجیستیک هستند، با هم مطابقت دارند؟

۳. آیا تحلیل محتوایی سوالات شامل DIF پیدا شده، نشان دهنده‌ی وجود سوگیری در آزمون سراسری کارشناسی ارشد رشته‌ی آموزش زبان انگلیسی است؟

روش پژوهش

داده‌های مورد نیاز برای مطالعه‌ی حاضر به صورت تصادفی از طریق پاسخنامه‌های ۲۹۵۲ نفر از داوطلبان ایرانی آزمون کنکور سراسری کارشناسی ارشد رشته‌ی آموزش زبان انگلیسی سال ۲۰۱۵ جمع‌آوری شد. برای رسیدن به اهداف این مطالعه، شرکت‌کنندگان به ترتیب بر اساس جنسیت به دو گروه تقسیم شدند: گروه مردان به عنوان گروه کانونی و گروه زنان به عنوان گروه مرجع در نظر گرفته شدند. گروه مردان شامل ۷۳۵ نفر (۸۹/۲۴ درصد) و گروه زنان شامل ۲۲۱۷ نفر (۷۵/۱۱ درصد) بودند. شرکت‌کنندگان آزمون کنکور سراسری کارشناسی ارشد رشته‌ی آموزش زبان انگلیسی کسانی بودند که مایل به ورود به دانشگاه‌های دولتی ایران بودند که در سطح کارشناسی ارشد در رشته‌ی آموزش زبان انگلیسی برنامه‌هایشان را ارائه می‌کنند. علاوه بر این، سه زبان‌شناس متخصص در این مطالعه کمک کردند و سوالات شامل DIF شناسایی شده در قسمت زبان عمومی‌آزمون را از لحاظ محتوایی تجزیه و تحلیل کردند تا دلایل احتمالی سوگیری سوالات شامل DIF مشخص شوند.

ابزار مورد استفاده در تحقیق حاضر شامل آزمون کنکور سراسری کارشناسی ارشد رشته‌ی آموزش زبان انگلیسی و پرسشنامه‌ی محتوایی است. اطلاعات آن‌ها به شرح زیر ارائه می‌شود:

آزمون کنکور سراسری کارشناسی ارشد رشته‌ی آموزش زبان انگلیسی آزمونی است که هر سال در ایران برگزار می‌شود تا دارندگان مدرک کارشناسی را که مایل به پذیرش در دانشکده‌های زبان انگلیسی دانشگاه‌های ایران می‌باشند مورد ارزیابی قرار دهد. این آزمون برای ارزیابی دانش زبان عمومی‌و تخصصی دانشجویانی که مایل به پذیرش در دوره کارشناسی ارشد رشته‌ی آموزش زبان انگلیسی در دانشگاه‌های ملی ایران هستند، طراحی شده است.

در مطالعه‌ی حاضر تنها بخش زبان عمومی‌آزمون که شامل ۶۰ سوال بود انتخاب شد. این بخش شامل سه قسمت بود: دستور زبان (۱۰ سوال)، واژگان (۲۰ سوال)، و درک مطلب (۲۰ سوال). همچنین، تعدادی سوال چهارگزینه‌ای شامل دستور زبان و واژگان در قالب یک آزمون کلوز (۱۰ مورد) وجود دارد.

این آزمون نمره‌ی منفی دارد. یعنی به ازای سه پاسخ نادرست یک پاسخ درست از بین می‌رود. به عبارت دیگر، اگر یک فرد شرکت کننده به سه سوال پاسخ نادرست دهد، یکی از پاسخ‌های صحیح وی حذف خواهد شد. در واقع، پاسخ ندادن به یک سوال تأثیری در نمره‌دهی آزمون نمی‌گذارد.

برای انجام بررسی محتوایی، پرسشنامه‌ای که توسط گرانپایه و کونان (۲۰۰۷) طراحی شده، مورد استفاده قرار گرفت. انتظار می‌رفت پرسشنامه تعیین کند که آیا آزمون کارشناسی ارشد آموزش زبان انگلیسی نسبت به جنسیت خاصی سوگیری داشته است یا خیر. پرسشنامه مقیاس ۵ تایی معیار لیکرت را شامل می‌شد که در آن ۱ (کاملا به نفع)، ۲ (به نفع)، ۳ (نه به نفع و نه به ضرر)، ۴ (به ضرر) و ۵ (کاملا به ضرر) را نشان می‌دهد. از کارشناسان زبان‌شناسی نیز خواسته شد تا شرح دهند که چرا برخی سوالات گروهی را بر دیگری مزیت داده یا به ضرر آن‌ها بوده است. به منظور کسب اطمینان از پایداری مقادیر تعیین شده، پایایی میان آزمونگر صورت گرفته توسط سه زبانشناس نیز محاسبه گردید.

سازمان سنجش آموزش کشور ایران داده‌های مورد نیاز برای این مطالعه را در اختیار قرار داده است. این سازمان دولتی هرساله مسئولیت برگزاری آزمون ورودی دانشگاه‌های

ایران را در مقاطع مختلف بر عهده دارد که کارشناسی ارشد یکی از آن‌ها است. داده‌ها بصورت یک فایل SPSS دریافت شده بود و محقق تغییراتی روی آن اعمال ننموده است. محقق، بعد از حصول داده، آمار توصیفی را با استفاده از SPSS نسخه ۲۲ محاسبه نمود. نمرات میانگین هر گروه، نمرات حداقل و حداکثر، خطای استاندار میانگین‌ها، و انحراف از معیار محاسبه گردید.

برای مقایسه‌ی میانگین گروه‌ها، آزمون تی (t-test) نمونه‌های مستقل محاسبه شد. برای یافتن موارد سوگیری نسبت به جنسیت خاص، شرکت کنندگان در آزمون بر مبنای جنسیت به گروه اصلی و مرجع دسته‌بندی گردیدند. شرکت کنندگان خانم (با کد ۱) گروه مرجع و شرکت کنندگان آقا (با کد ۲) گروه اصلی تعیین شدند. داده‌های این مطالعه دوجهی است و به هر پاسخ مثبت ۱ و هر پاسخ غلط ۰ تعلق می‌گیرد. آزمون دهندگان براساس نمره‌ی کلی در آزمون زبان عمومی، با یک‌دیگر تطبیق داده شدند. سپس تحقیق به بررسی چگونگی عملکرد گروه‌های مختلف در هر سوال پرداخت و سوالاتی که رفتار متفاوتی از خود نشان می‌دادند را شناسایی نمود.

برای شناسایی سوالات شامل DIF، منتل-هنسزل برای هر گروه محاسبه گردید. آمار منتل-هنسزل را می‌توان توسط نرم‌افزار SPSSسهل‌الوصول و یا بسته‌ی تخصصی‌تر (DIFAS: پنفیلد[1] ۲۰۰۵، ۲۰۱۰) محاسبه کرد. در این مطالعه برای یافتن DIF منتل-هنسزل از DIFA ۵,۰ (پنفیلد، ۲۰۱۰) استفاده شد. با هدف بررسی داده‌ها از زاویه‌ی دیگر، رگرسیون لجیستیک به عنوان روش دوم آنالیزDIF مورد استفاده قرار گرفت. برای محاسبه‌ی رگرسیون لجیستیک در داده‌های اسمی، دستور SPSS ناگلکرک[2] (زامبو[3]، ۱۹۹۹) استفاده شد. سپس یافته‌ها با نتایج منتل-هنسزل مقایسه شد تا پایایی یافته‌های هر گروه از آزمون دهندگان تایید گردد. جزییات روش‌های بررسیDIF و مراحل آن در بخش بعد آمده است.

در این تحقیق از دو روش اصلی آنالیز استفاده شد: آنالیز آماری داده و آنالیز محتوایی داده. آنالیز آماری به وسیله SPSS ۲۲ انجام شد که به بررسی آمار توصیفی، آزمون تی نمونه‌های مستقل، و همبستگی پیرسون پرداخت. به علاوه روش‌های منتل-هنسزل (MH) و

1. Penfield
2. Nagelkerke's SPSS
3. Zumbo

رگرسیون لجیستیک (LR) برای شناسایی سوال‌های دارای DIF آزمون ورودی کارشناسی ارشد آموزش زبان انگلیسی استفاده شد. آنالیز محتوایی داده نیز توسط سه متخصص زبان-شناسی انجام شد که درباره سوگیری سوالات به نفع گروه خاصی از شرکت‌کنندگان نظراتی ارائه کردند.

نتایج پژوهش

اطلاعاتی که از پاسخنامه‌ی ۲۹۵۲ نفر از آزمون دهندگان کارشناسی ارشد رشته‌ی TEFL به دست آمد، نشان داد که نمره‌ی میانگین این آزمون ۶۰ سوالی، ۱۱/۶۶ بود که انحراف معیار آن ۵/۸۰ به دست آمد. جدول ۱ برخی از آمارهای توصیفی این آزمون دهندگان را بدون در نظر گرفتن جنسیت آن‌ها خلاصه می‌کند.

جدول ۱. آمارهای توصیفی

	تعداد کل	حداقل	حداکثر	میانگین	خطای استاندارد میانگین‌ها	انحراف معیار
نمره	۲۹۵۲	۰/۰۰	۴۵/۰۰	۱۱/۶۶	۰/۱۰	۵/۸۰

همان‌طور که در جدول ۱ دیده می‌شود، بیشترین نمره‌ی کسب شده توسط آزمون‌دهندگان ۴۵ از ۶۰ پاسخ درست بوده و کمترین نمره صفر بوده است، زیرا برای جواب‌های نادرست نمره‌ی منفی در نظر گرفته نشده بود.

تنها متغیری که در مطالعه‌ی سوال‌های جدا شده‌ی این تحقیق مورد بررسی قرار گرفت، جنسیت بود. از میان ۲۹۵۳ آزمون دهنده، ۷۳۵ نفر مرد و ۲۲۱۷ نفر زن بودند. میانگین نمره‌ی آزمون‌دهندگان آقا ۱۲/۲۴ و میانگین نمره‌ی آزمون دهندگان خانم ۱۱/۴۷ بود. جدول ۲ بخشی از آمار توصیفی گروه جنسیت را نشان می‌دهد.

جدول ۲. آمارهای گروه جنسیت

جنسیت	تعداد	نمره میانگین	انحراف از معیار	خطای‌استاندارد میانگین‌ها
زن	۲۲۱۷	۱۱/۴۷	۵/۶۲	۰/۱۱
مرد	۷۳۵	۱۲/۲۴	۶/۲۹	۰/۲۳

جدول ۲ به وضوح نشان می‌دهد که انحراف معیار محاسبه شده برای آزمون‌دهندگان مرد و زن، به ترتیب ۶/۲۹ و ۵/۶۲ است. تفاوت اندک میانگین (۰/۶۷=۵/۶۲ -۶/۲۹) نشان‌دهنده‌ی تغییرپذیری پایین است، هرچند برای مشخص کردن این که آقایان و خانم‌ها

در توانایی پنهان مورد آزمون به شکل معناداری تفاوت دارند یا خیر، یک آزمون تی نمونه‌های مستقل اجرا شد. جدول ۳ نتایج به دست آمده از آزمون تی را نشان می‌دهد.

جدول ۳. آزمون تی نمونه‌های مستقل برای جنسیت

		آزمون لون برای برابری واریانس‌ها		آزمون تی برای برابری میانگین‌ها		
		F	Sig.	t	Df	Sig. (2-tailed)
cScore	Equal variances assumed	۱۶/۰۱	۰/۰۰	-۳/۱۲۲۹۵۰		۰/۰۰۲

توجه: معنادار در$p<$ ۰/۰۵

نتایج آزمون تی نمونه‌های مستقل که در جدول ۳ نشان داده شده است، نشان‌دهنده‌ی تفاوتی قابل توجه میان نمره‌ی میانگین آزمون دهندگان زن و مرد است (۰/۰۰۱= p، ۳/۱۲- = (۲۹۵۰) t). با این که به نظر می‌آید گروه مردان نمره‌ی میانگین بالاتری نسبت به گروه زنان دارند، اندازه اثر چنین تفاوت اندکی بین میانگین‌ها (۰/۷۷) ۰/۱۳ است که تأثیر اندازه‌ی کمی‌محسوب می‌شود.

کوه[1] (۱۹۸۸) برای اندازه‌گیری اندازه اثر روشی را پیشنهاد داد. او معیاری برای اندازه اثر معرفی کرد که با استفاده از محسابه‌ی تفریق میان میانگین‌ها تقسیم بر انحراف از معیار گروه‌ها به دست می‌آید. کوهن (۱۹۸۸) میان تأثیر اندازه‌های کوچک، متوسط و بزرگ تمایز قائل شد. طبق نظر وی ۰/۱۳ اندازه اثر کم تلقی می‌شود و تفاوت نمره‌ی میانگین بین آزمون دهندگان آقا و خانم قابل چشم پوشی است. بنابراین این دو گروه قابل مقایسه هستند.

نتایج تحلیل منتل-هنسزل برای خانم‌ها وآقایان در جدول ۴ نشان داده شده است. فقط سوالاتی که دارای DIF هستند در این جدول قرار گرفته‌اند. برنامه‌ی دیفاز (پنفیلد، ۲۰۰۹) با استفاده از تمام مراحل تشخیصی که در بالا به آن اشاره شد در مطالعه‌ی فعلی به کار رفت تا سوالات دارای DIF جنسیتی پیدا شوند. از بین ۶۰ سوال انگلیسی عمومی‌در آزمون، ۱۱ سوال (۱۸/۳۳٪) کنش افتراقی نسبت به آزمون‌دهندگان مرد یا زن داشتند. یعنی به عنوان سوالاتی دارای DIF جنسیتی منتل-هنسزل شناخته شدند. با توجه به طرح طبقه‌بندی ETS، ۹ سوال در بخش A (قابل چشم‌پوشی) اندازه اثر[2] جای گرفتند و فقط ۱ سوال در بخش B

1. Cohen
2. Effect Size

(متوسط) DIF قرار گرفت. همچنین یک سوال در بخش C (بزرگ(DIF قرار گرفت. جدول ۴ نتایج به دست آمده را نشان می‌دهد.

جدول ٤. جدول منتل-هنسزل برای جنسیت

شماره سوال	MH CHI	MH LOR	LOR SE	LOR Z	BD	CDR	ETS
۱	۹/۶۱	۰/۳	۰/۰۹	۳/۱۱	۲/۲۶	Flag	A
۳	۷/۲۰	۰/۲۸	۰/۱۰	۲/۷۲	۰/۱۴	Flag	A
۵	۷/۰۰	۰/۲۵	۰/۰۹	۲/۶۸	۱/۵۳	Flag	A
۲۰	۷/۱۷	-۰/۳۹	۰/۱۶	-۲/۳۸	۰/۴۱	Flag	A
۲۷	۴۳/۹۳	-۰/۸۰	۰/۱۲	-۶/۵۵	۰/۱۶	Flag	C
۲۸	۱۳/۰۳	-۰/۶۰	۰/۱۶	-۳/۶۸	۱/۶۱	Flag	B
۳۰	۵/۴۷	-۰/۲۷	۰/۱۲	-۲/۴۰	۰/۰۰	Flag	A
۴۵	۱۲/۷۰	-۰/۳۹	۰/۱۰	-۳/۶۱	۰/۰۰	Flag	A
۴۶	۱۲/۶۱	۰/۳۳	۰/۰۹	۳/۵۹	۰/۰۷	Flag	A
۴۸	۰/۲۵	-۰/۰۷	۰/۱۲	-۰/۵۶	۶/۶۵	Flag	A
۵۵	۷/۳۷	۰/۳۴	۰/۱۲	۲/۷۵	۴/۷۷	Flag	A

توجه: MH CHI: مجذور فی منتل-هنسزل؛MH-LOR: نسبت شانس لگاریتمیMH؛ LOR SE: نسبت خطای استاندارد منتل-هنسزل LOR Z: نسبت استانداردشده‌ی شانس لگاریتمی MH BD: مجذور فی برسلو-دی CDR: قانون ترکیبی تصمیم‌گیری ETS: طرح طبقه بندی ای تی اس

نتایج نشان می‌دهند که تمام سوالات ساختاری دارای DIF به نفع خانم‌ها بود (۱، ۳ و۵). مقادیر مثبت آمارهای MH-LOR (که سوگیری به سوی گروه مرجع دارد) نشان‌دهنده‌ی این سوگیری است.

بر خلاف سوالات ساختاری، تمام چهار سوال واژگان، به نفع آقایان سوگیری داشتند (۲۰، ۲۷، ۲۸ و ۳۰). می‌توان این موضوع را با مشاهده‌ی آمار MH LOR در ستون دوم جدول ۴ بهتر دید. مقادیر این کنش آماری منفی هستند و سوگیری به سمت گروه کانونی را نشان می‌دهند.

چهار سوال بخش درک مطلب هم دارای کنش افتراقی نسبت به گروه‌های جنسی بودند. این سوالات شامل سوال‌های ۴۵، ۴۶، ۴۸ و۵۵ هستند. سوال‌های ۴۵ و۴۸، با مقادیر منفی

MH LOR دارای سوگیری به سمت گروه آقایان بودند و دو سوال دیگر دارای DIF، یعنی سوال ۴۶ و ۵۵، با مقادیر مثبت MH LOR به سمت گروه خانم‌ها سوگیری داشتند.

با استفاده از روش رگرسیون لجیستیک، سوالات آزمون کارشناسی ارشد رشته‌ی آموزش زبان انگلیسی (۶۰ سوال) تک به تک تحلیل شد و نتایج آن در جدول ۵ نشان داده شده است. در این جدول، تنها سوالاتی که دارای سطح معنی‌دار DIF بودند گنجانده شده‌اند (در این پژوهش از سطح معنی‌دار ۰/۰۱ استفاده شده است). به نظر می‌رسد که آزمون مجذور فی (کای اسکوئر) با دو درجه‌ی آزادی در ۸ سوال برای DIF دارای سطح معنی‌دار بود. از این ۸ سوال یک سوال در قسمت گرامر (سوال ۱)، دو سوال در قسمت لغات (سوالات ۲۷ و ۲۸) و پنج سوال در قسمت درک مطلب (سوالات ۴۵، ۴۶، ۴۸، ۵۴، و ۵۵) یافت شدند. در واقع در قسمت آزمون کلوز هیچ سوال شامل DIF یافت نشد.

جدول ۵. نتایج DIF جنسیتی رگرسیون لجیستیک

شماره سوال	آزمون مجذور فی DIF	گام ۲ – گام ۱ $DIF\ R^2$ هماهنگ	گام ۳ – گام ۲ $DIF\ R^2$ ناهماهنگ	گام ۳ – گام ۱ اندازه ی $DIF\ R^2$	هماهنگ /ناهماهنگ
۱	۱۱/۵۴	۱۰/۴۱	۱/۱۳	۰/۰۰	هماهنگ
۲۷	۴۳/۵۸	۴۱/۹۰	۱/۶۸	۰/۰۲	ناهماهنگ
۲۸	۷۱,۱۷	۱۵/۸۶	۱/۸۵	۰/۰۱	ناهماهنگ
۴۵	۱۵/۹۲	۸/۲۸	۵/۰۹	۰/۰۰	هیچکدام
۴۶	۲۲/۹۴	۱۷/۷۳	۵/۲۰	۰/۰۰	هماهنگ
۴۸	۲۴/۰۶	۰/۷۳	۲۳/۹۴	۰/۰۱	ناهماهنگ
۵۴	۱۲/۵۵	۵/۸۷	۶/۶۷	۰/۰۰	هیچکدام
۵۵	۳۲/۰۷	۱۲/۳۳	۱۹/۷۴	۰/۰۱	هر دو

بر اساس پیشنهادات سوامیناتان و راجرز (۱۹۹۰) این آزمون با استفاده از بدست آوردن مقدار مجذور فی برای گام ۳ منهای مقدار مجذور فی گام ۱ انجام شد. مقدار مجذور فی بدست آمده در جدول با کنش توزیع مربوطه‌اش با دو درجه‌ی آزادی مقایسه شد. در صورتیکه آن مقدار دارای سطح معنی‌داری باشد آن سوال به عنوان یک سوال شامل DIF در نظر گرفته می‌شود (مقداری برابر یا بیشتر از ناحیه‌ی بحرانی مقدار احتمال ۰/۰۱ که ۹/۲۱ است). ستون دوم جدول این مقادیر را نشان می‌دهد. برای بدست آوردن DIF هماهنگ،

مقدار مجذور فی برای گام ۱ از مقدار مجذور فی گام ۲ کم شد (ستون سوم جدول). DIF ناهماهنگ با تفریق مقدار مجذور فی گام ۲ از مقدار مجذور فی گام ۳ به دست آمد (ستون چهارم جدول).

در این مطالعه، مقادیر مجذور فی به دست آمده نشان دادند که نیمی‌از سوالات DIF (چهار سوال) به عنوان سوالات DIF هماهنگ شناخته شدند: سوالات ۱، ۲۷، ۲۸، و ۴۶. فقط یک سوال، سوال ۴۸ به عنوان سوالات DIF ناهماهنگ شناخته شد. در کنار این نتایج چند نتیجه‌ی جالب هم به دست آمد. دو سوال DIF (سوالات ۴۵ و ۵۴) نه به عنوان DIF هماهنگ و نه به عنوان DIF ناهماهنگ شناخته شدند. برخلاف آن‌ها، سوال ۵۵ هم به عنوان DIF هماهنگ و هم به عنوان DIF ناهماهنگ نشان داده شد. درمورد اندازه‌ی تاثیر DIF بر اساس طبقه‌بندی محافظه‌کارانه‌ی جدوین و گیرل (۲۰۰۱) این نتیجه حاصل شد که تمامی-سوالات DIF که با روش رگرسیون لجیستیک پیدا شده بودند در طبقه ی A قرار دارند. این یعنی تمامی‌سوالات DIF پیداشده با روش رگرسیون لجیستیک اندازه‌ی تاثیر ناچیز DIF را نشان دادند. این نتایج بر اساس راهنماهای زامبو و توماس[1] (۱۹۹۷) و جدوین و گیرل[2] (۲۰۰۱) به دست آمد.

درمورد سوگیری سوالات DIF پیداشده، نتایج نشان داد که از ۸ سوال DIF، ۴ سوال به نفع شرکت‌کنندگان زن سوگیری داشتند و ۴ سوال هم به نفع شرکت‌کنندگان مرد متمایل بودند. سوالاتی که به نفع زنان گرایش داشتند سوالات ۱ (ساختار)، ۴۶، ۵۴، و ۵۵ (درک مطلب) بودند. همان‌طور که مشاهده می‌کنید سوالات درک مطلب اساساً به نفع شرکت-کنندگان خانم سوگیری داشتند. اما هیچ‌کدام از سوالات لغات به آنان گرایش نداشتند و در عوض، سوگیریشان به سمت گروه شرکت کنندگان مرد بوده است (سوالات ۲۷ و ۲۸).

با توجه به سوال دوم این تحقیق، درمورد امکان مقایسه‌ی نتایج دو روش منتل- هنسزل و رگرسیون لجیستیک، این نتیجه به دست آمد که با استفاده از روش منتل- هنسزل ۱۱ سوال، نشان دهنده ی DIF بودند و با استفاده از روش رگرسیون لجیستیک ۸ سوال شامل DIF تشخیص داده شدند. جدول ۶ مقایسه‌ی بین نتایج این دو روش آماری را نشان می‌دهد.

جدول ٦. مقایسه ی نتایج روش های منتل- هنسزل و رگرسیون لجیستیک

1. Zumbo & Thomas
2. Jodoin & Gierl

آزمون	قسمت	نتایج رگرسیون لجیستیک (اندازه تاثیر)	نتایج منتل-هنسزل (اندازه تاثیر)	شماره سوال
ساختار		LR(A)	MH (A)	۱
ساختار		-	MH (A)	۳
ساختار		-	MH (A)	۵
لغات		-	MH (A)	۲۰
لغات		LR (A)	MH (C)	۲۷
لغات		LR (A)	MH (B)	۲۸
لغات		-	MH (A)	۳۰
درک مطلب		LR (A)	MH (A)	۴۵
درک مطلب		LR (A)	MH (A)	۴۶
درک مطلب		LR (A)	MH (A)	۴۸
درک مطلب		LR (A)	–	۵۴
درکمطلب		LR (A)	MH (A)	۵۵

توجه: MH: منتل-هنسزل؛ LR: رگرسیون لجیستیک؛ A: اندازه تاثیر ناچیز DIF؛ B: اندازه تاثیر متوسط DIF؛ C: اندازه تاثیر قوی DIF

همان‌طور که در جدول مشاهده می‌کنید، با استفاده از روش منتل-هنسزل، ۳ سوال دارای DIF بیشتر از روش رگرسیون لجیستیک پیدا شد. این یافته‌ها نتایجی متناقض با یافته‌های سوامیناتان و راجرز (۱۹۹۰) نشان می‌دهد که مدعی بودند فرآیند رگرسیون لجیستیک به اندازه‌ی فرآیند منتل- هنسزل در پیدا کردن DIF هماهنگ قدرت دارد و در پیدا کردن DIF ناهماهنگ از روش منتل-هنسزل قوی‌تر است. این مطالعه ثابت کرد که فرآیند منتل-هنسزل از فرآیند لجستیک رگرسیون در یافتن DIF هماهنگ قوی‌تر عمل می‌کند.

همان‌طور که محققانی مثل کامیلی و شپارد (۱۹۹۴) و آنگوف (۱۹۹۳) بیان کرده‌اند، برای این که یک سوال سوگیر شناخته شود، داشتن DIF ضروری است اما کافی نیست. به عبارت دیگر ممکن است یک سوال دارای کنش افتراقی جنسیتی شناخته شود، اما در واقع نسبت به گروه خاصی از آزمون دهندگان سوگیری نداشته باشد. تحلیل‌های بیشتری در محتوای این سوال‌ها می‌تواند کمک‌مان کند تا منبع احتمالی DIF را پیدا کنیم.

هرچند میزان مطالعه برای تحلیل محتوای سوال‌های دارای DIF بسیار کم است (اوترویجک و والن[1]، ۲۰۰۵)، اشمیت، هلند، و دورانز (۱۹۹۲) برخی از دلایل چنین بی توجهی به منابع DIF را توضیح داده‌اند. تازگی این موضوع و پیچیدگی یافتن دلایل DIF برخی از این دلایل بود. با این حال گاهی نمی‌توان دلیل واقعی یک سوال دارای DIF را پیدا کرد.

به منظور پاسخگویی به سومین سوال تحقیق و یافتن منابع زبان‌شناختی برای سوال‌های دارای کنش افتراقی جنسیتی، با کمک سه متخصص زبان‌شناسی یک آنالیز محتوایی انجام شد. بدین منظور، پرسشنامه‌ی ۵ بخشی گرانپایه و کونان[2] (۲۰۰۷) مورد استفاده قرار گرفت. زبان‌شناسان بر سوگیری جنسیتی سوال‌های دارای کنش افتراقی جنسیتی نظر دادند و دلایلشان را مطرح نمودند. این پرسشنامه درمورد منافع و یا مضرات سوالات حاویDIF برای گروه‌های جنسیتی مرد و زن بود. بر اساس یافته‌های تحلیل محتوایی متخصصان زبان-شناسی، هیچ منبع زبانی خاصی برای سوالات حاوی DIF پیدا نشد. درواقع، دو سوال از بخش لغات (سوالات ۲۷ و ۲۸) از نظر کارشناسان زبان‌شناسی دارای دلایلی برای سوگیری متفاوتشان نسبت به گروه‌های جنسیتی بودند، ولی آن دلایل به فرایندهای روان‌شناسی و ذهنی مرتبط می‌شدند. از نظر کارشناسان زبان‌شناسی، سوالات ۲۷ و ۲۸ متعلق به بخش لغات، به‌خاطر سختی و پیچیدگی واژگان به نفع آقایان سوگیری دارند. آن‌ها مدعی شدند که آزمون دهندگان آقا اغلب واژگان بیشتری از خانم‌ها می‌دانند (مخصوصاً واژگان پیچیده). هیچ منبع زبان‌شناسی خاصی برای علت سوگیری سوالات شامل DIF گزارش نشده است.

هیچ زبان‌شناسی درباره‌ی سوال گرامری شامل DIF که در هر دو روش منتل-هنسزل و رگرسیون لجیستیک پیدا شد، دلیلی برای سوگیری این سوال عنوان نکرد. برای سوال شامل DIF در درک مطلب، هیچ منبع زبانشناختی مبنی بر سوگیری جنسیتی در سوال توسط زبان-شناسان مشخص نشد.

به طور خلاصه، یافته‌های روش رگرسیون لجستیک نشان داد که ۸ سوال در آزمون بین جنسیت‌ها سو گیری ایجاد می‌کند. از ۶۰ سوال بخش زبان عمومی‌آزمون ۱۳/۳۳ درصد سوگیری جنسیتی دارند. روش MH نشان می‌دهد که ۴۵/۴۵ درصد سوالات دارای DIF،

1. Uiterwijk & Vallen
2. Geranpayeh & Kunnan

به نفع شرکت‌کنندگان خانم و ۵۴/۵۵ درصد دیگر به نفع شرکت‌کنندگان آقا عمل می‌کنند. این انتظار به وجود آمد که شرکت‌کنندگان آقا بهتر از خانم‌ها عمل کنند و یافته‌های پژوهش این موضوع را تصدیق کرد.

مطالعات پیشین در خصوص سوگیری جنسیتی، نتایج متفاوتی را به دنبال داشته است. مثلا ریان و بکمن (۱۹۹۲) تحقیقی در مورد آزمون های FCE و تافل انجام دادند و هیچ سوال دارای DIF جنسیتی پیدا نکردند. لین و وو (۲۰۰۳) نیز در مطالعه ی خود بر کنش افتراقی سوالات آزمون EPT در چین، هیچ سوالی نیافتند که گروه جنسیتی خاصی را هدف بگیرد. از میان محققان ایرانی که در حوزه ی DIF جنسیتی مطالعه می‌کردند، امیریان (۲۰۱۲) در بررسی خود از آزمونUTEPT نتایج متناقضی با آنچه که در این مطالعه ارائه شده یافتند. نتایج آن تحقیق نشان داد کهUTEPT بیشتر به نفع خانم‌ها است. هچنین کرمی-(۲۰۱۱) مطالعه‌ی DIF روی UTEPT انجام داد که دریافت این آزمون برای شرکت-کنندگان خانم ساده تر بوده و به ضرر آقایان عمل کرده است.

از میان ۶ سوالی که به نفع آقایان بود ۴ مورد به بخش واژگان مربوط می‌شد. به بیان دیگر تمامی‌سوالات واژگان دارای DIF برای آقایان ساده تر بود. جالب توجه این که سه سوال دستور زبان کهDIF داشته به نفع خانم‌ها بوده است و این می‌تواند به این نتیجه منجر شود که گرامر و واژگان بطور قابل توجهی بین خانم‌ها و آقایان تفاوت ایجاد می‌کنند. بنابراین عامل تفاوت در کنش این دو گروه جنسیتی می‌باشد. جالب است که هم امیریان (۲۰۱۲) و هم کرمی(۲۰۱۱) به همین نتیجه دست یافته‌اند. همچنین در آزمون UTEPT براساس آنچه که آن‌ها گزارش می‌کنند بخش گرامر به نفع شرکت‌کنندگان خانم و واژگان به نفع آقایان بوده است. البته نتایج گزارش شده توسط رضایی و شعبانی (۱۳۸۸) متفاوت بوده است. در واقع آن‌ها بخش درک مطلب را عامل تفاوت در کنش آقایان و خانم‌ها به شمار آورده‌اند. البته در این مطالعه ۵ مورد DIF نیز در بخش درک مطلب آزمون کارشناسی ارشد بوده است که میان سوالاتی که به نفع آقایان و خانم‌ها بوده تعادل ایجاد کرده است زیرا سه مورد به نفع خانم‌ها و دو مورد به نفع آقایان بوده است. در بخش درک مطلب تافل که توسط وینر و لوخل (۱۹۹۷) بررسی شده بود هیچ تمایلی به سمت گروه جنسیتی خاصی گزارش نشد. در رابطه با سومین سوال مطرح شده در این تحقیق، هیچ منبع زبان‌شناسی برای DIF موارد یافت شده گزارش نشده است. یافته‌های این پژوهش با مطالعه‌ی انجام شده

توسط کونان و وینستر-شر (۱۹۹۰) منطبق است زیرا آن‌ها نیز موفق نشده بودند تا یافتن علت سوگیری ۱۱ مورد سوالات شامل DIF بخش واژگان و گرامر را از نظر زبان‌شناسی بیابند.

بحث و نتیجه‌گیری

هدف مطالعه‌ی DIF حاضر، یافتن سوالاتی بود که برای گروه خاصی از شرکت‌کنندگان خانم یا آقا به صورت متفاوتی عمل می‌کردند. همچنین تلاش‌هایی در جهت یافتن ریشه‌ی زبان‌شناختی سوگیری سوالات شامل DIF یافته شده صورت پذیرفت. نتایج حاکی از آن بود که سوالات آزمون کارشناسی ارشد آموزش زبان انگلیسی، به جز دو مورد، نسبت به شرکت‌کنندگان مرد و زن سو گیری نداشته است. به بیان دیگر آزمون کارشناسی ارشد رشته‌ی آموزش زبان انگلیسی در برابر گروه‌های جنسیتی مختلف عادلانه و بی‌طرفانه بوده است. در نهایت، بررسی محتوایی انجام گرفته توسط سه زبان‌شناس، از منظر زبان‌شناسی منبع خاصی را برای نمونه‌های DIF پیدا نکرد.

از یافته‌های به‌دست آمده در این مطالعه این نتیجه گیری حاصل می‌شود که جنسیت نقش اساسی در پاسخگویی افراد به سوالات هر آزمونی ایفا می‌کند و می‌تواند باعث تفاوت پاسخگویی آزمون دهندگان آزمون‌های سرنوشت ساز و بزرگ شود. سوگیری سوالات به جنسیت جزو عوامل نامربوط موثر در آزمون به شمار می‌رود و باعث می‌شود آزمون فاقد عدالت باشد. پس چنین عوامل نامربوطی که هدف ارزیابی سوالات نیستند باید شناسایی و از سوالات آزمون‌ها حذف شوند.

درمورد کاربردهای این مطالعه می‌توان به این نکته اشاره کرد که طراحان آزمون و آزمون دهندگان آزمون‌های سراسری و کنکور باید توجه ویژه‌ای به عدالت آزمون داشته باشند و دقت کنند که سوالات این آزمون فاقد هر گونه سوگیری به نفع گروه خاصی از شرکت‌کنندگان باشند. با توجه به اینکه سوالات بخش لغات این آزمون برای شرکت‌کنندگان خانم سخت تر بودند، آموزگاران باید در تدریس خود به این نکته توجه کنند و به زبان‌آموزان خانم کمک کنند تا در یادگیری لغات پیچیده انگلیسی موفق باشند. همچنین در خصوص بخش سوالات گرامر مشخص شد که مردان به دلیل دید کلی‌ترشان نسبت به زنان ضعیف‌تر عمل کردند. بنابراین آموزگاران می‌توانند با ارائه‌ی تمرین‌های

گرامری جزئی‌تر به آموزندگان آقا به آن‌ها توجه ویژه‌ای در تدریس گرامر زبان انگلیسی داشته باشند.

از محدودیت‌های مطالعه‌ی حاضر می‌توان به مسئله‌ی تعداد نمونه‌های مورد مطالعه اشاره کرد. داده‌های تحقیق از پاسخنامه‌های تعداد ۲۹۵۲ شرکت کننده جمع‌آوری شده است. با اینکه این تعداد به اندازه‌ی کافی زیاد بود که نتایج قابل قبولی به دست آید ولی باید توجه داشت که روش منتل-هنسزل برای قدرت بیشتر در تشخیص کنش افتراقی به تعداد نمونه‌های بسیار زیادتری نیازمند است.

در این بخش برای محققان در زمینهDIF پیشنهاداتی ارائه گردیده است. به طور مثال فرد می‌تواند از روش‌های ردیابی DIF دیگری به جز MH و LR استفاده کند. روش‌هایی مانند نظریه پاسخ سوال (IRT) شامل مدل‌های رش (Rasch)، تحلیل واریانس (ANOVA) و تحلیل کوواریانس (ANCOVA)، روش‌های اریا (Area) ، روش‌های استانداردسازی، روش مجذور فی لرد، روش سیب تست، آزمون نسبت مشابهت IRT و دیگر روش‌های این چنینی برای یافتن سوالات شامل DIFجنسیتی قابل استفاده هستند. علاوه بر جنسیت فرد، عوامل بی ارتباط با ساختار، از جمله سابقه‌ی تحصیلی ، سن، و یا تعلق داشتن به قومیت خاص می‌توانند موضوعات خوبی برای مطالعه‌یDIF باشند. آزمون کارشناسی ارشد رشته‌ی آموزش زبان انگلیسی بر آینده‌ی شرکت‌کنندگان تأثیر به سزایی دارد. بنابراین باید از دیدگاه‌های مختلفی مورد بررسی دقیق قرار بگیرد. این مطالعه، به بررسی نحوه کنش آزمون‌دهندگان پرداخت تا سوگیری آزمون مشخص گردد. برای بررسی اعتبار و عدالت آزمون محققان می‌توانند از روش‌های دیگری مانند بررسی سوالات آزمون به جای کنش شرکت‌کنندگان استفاده نمایند.

منابع

خدایی، ابراهیم. (۱۳۸۸). بررسی رابطه سرمایه اقتصادی و فرهنگی والدین دانش آموزان با احتمال قبولی آنها در آزمون سراسری سال تحصیلی ۱۳۸۵. *فصلنامه انجمن آموزش عالی ایران*، ۴(۱)، ۶۵-۸۴

رضایی، عباسعلی.، و شعبانی، عنایت‌اله. (۱۳۸۸) تحلیل کارکرد افتراقی جنسیتی آزمون سنجش توانش عمومی زبان دانشگاه تهران[ویژه نامه]. *پژوهش زبان‌های خارجی، ۵۶*، ۸۹-۱۰۸.

نوربخش، سید مرتضی. (۱۳۸۹). نقش سرمایه های فرهنگی، اجتماعی و اقتصادی خانواده در موفقیت داوطلبان آزمون سراسری. *فصلنامه برنامه‌ریزی رفاه و توسعه اجتماعی*، ۴(۱)، ۹۳-۱۳۴.

Alavi, S. M., & Karami, H. (2010). Differential item functioning and ad hoc interpretations. *TELL*, *4*(1), 1-18.

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*, *36*(3), 185-198.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC.

Amirian, S. M. R. (2012). *Investigating UTEPT for gender and academic discipline*. (Unpublished PhD dissertation). Tehran University, Tehran.

Angoff, W. H. (1993). Perspectives on Differential Item Functioning methodology. In P. W. Holland, & H. Wainer (Eds.), *Differential Item Functioning* (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum.

Birjandi, P. & Amini, M. (2007). Differential item functioning: The case of Mantel-Haenszel and functioning (Test Bias) analysis paradigm across Manifest and Latent examinee groups (on the construct validity of IELTS). *Human Sciences*, *8*(2) 1-20.

Camilli, G. (2006). Test fairness. In R. Brennan (Eds.), *Educational measurement* (pp. 221-256). Westport, CT: American Council on Education and Praeger.

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

Carlton, S. T., & Harris, A. M. (1992). *Characteristics associated with Differential Item Functioning on the Scholastic Aptitude Test: Gender and majority/minority group comparisons* (Report no. 64). Princeton, NJ: Educational Testing Service.

Cohen, A. S. (1988). *Statistical power analysis for behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Association.

Conoley, C. A. (2003). *Differential item functioning in the Peabody Picture Vocabulary Test–Third Edition: Partial correlation versus Expert judgment.* (Unpublished doctoral dissertation). Texas A&M University, TX.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting Differential Item Functioning in polytomous items. *Journal of Educational Measurement*, *33*(3), 315-332.

Geranpayeh, A., & Kunnan, A. J. (2007). Differential Item Functioning in terms of age in the certificate in advanced English examination. *Language Assessment Quarterly*, *4*(2), 190-222.

Holland, P. W. & Thayer, D. T. (1988). Differential item performance and Mantel- Haenszel procedure. In H. Wainer & H. Braun, (Eds.), *Test validity* (pp. 129-45). Hillsdale, NJ: Lawrence Erlbaum.

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

Karami, H. (2011). Detecting gender bias in a language proficiency test. *International Journal of Language Studies*, *5*(2), 27-38.

Kunnan, A. J. & Weinstein-SHR, G. (1990), DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly*, *24*(4), 741-746.

Lawrence, I. M., & Curley, W. E. (1989). *Differential item functioning for males and females on SAT-Verbal Reading sub-score items: Follow-up study*. Educational Testing Service (Report no. 22). Princeton, NJ: ETS.

Lawrence, I. M., Curley, W. E. & McHale, F. J. (1988). *Differential item functioning for males and females on SAT verbal reading subscore items.* (Report No. 88-4). New York: College Entrance Examination Board.

Lin, J., & Wu, F. (2003, April). *Differential performance by gender in foreign language testing*. Poster for the 2003 annual meeting of NCME in Chicago.

McNamara, T. F., & Roever, C. (2006). *Language testing: The social dimension*. Oxford, UK: Blackwell Publishing.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*(2), 105-118.

Pae, T. (2004). DIF for learners with different academic backgrounds. *Language Testing*, *21*(1), 53-73.

Penfield, R. D. (2010). Modeling DIF effects using distractor-level invariance effects: Implications for understanding the causes of DIF. *Applied Psychological Measurement*, *34*(3), 151-165.

Penfield, R. D. (2009). *Differential Item Functioning Analytical System*. DIFAS 5.0. User's manual.

Penfield, R. D. (2005). DIFAS: Differential item functioning analysis system. *Applied Psychological Measurement, 29*(2), 150-151.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement*, *20*(4), 355-71.

Ryan, K., & Bachman, L. F. (1992). Differential item functioning on two tests of EFL Proficiency. *Language Testing*, *9*(1) 12-29.

Schmitt, A. P., Holland, P. W., & Dorans, A. J. (1992). *Evaluating hypothesis about differential item functioning*. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 281-315). Hillsdale, NJ: Lawrence Erlbaum Associates.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal for Educational Measurement, 27*(4), 361-370.

Uiterwijk, H., & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. *Language Testing*, *22*(2), 211-234.

Wainer, H., & Luckele, R. (1997). How reliable are TOEFL scores? *Educational and Psychological Measurement, 57*(5), 741-759.

Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF* (Working paper of the Edgeworth Laboratory for Quantitative Behavioral Science). Prince George, Canada: University of Northern British Columbia.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.