



## پژوهشهای زبانشناختی در زبانهای خارجی

شاپای الکترونیکی: ۷۵۲۱-۲۵۸۸

شاپای چاپی: ۴۱۲۳-۲۵۸۸

وبسایت نشریه: [www.jflr.ut.ac.ir](http://www.jflr.ut.ac.ir)



### بررسی عملکرد افتراقی آیتم ها در آزمون ورودی دانشگاه با استفاده از آنالیزمدل راش

سوده بردبار

دکتری آموزش زبان انگلیسی،  
دانشگاه علوم پزشکی ایران، تهران،  
ایران  
(نویسنده مسئول)



سوده بردبار دارای مدرک کارشناسی ارشد و دکتری در رشته ی آموزش زبان انگلیسی از دانشگاه تهران در سال های ۲۰۰۹ و ۲۰۱۷ می باشد.

Email: [ut.sbordbar@yahoo.com](mailto:ut.sbordbar@yahoo.com)

چکیده	اطلاعات مقاله
پژوهش حاضر به منظور بررسی عملکرد افتراقی آیتم ها در سوالات آزمون ورودی دانشگاه های ایران در رشته زبان های خارجی (NUEEFL) با در نظر گرفتن جنسیت شرکت کنندگان انجام شده است. شرکت کنندگان (تعداد = ۵۰۰۰) در این مطالعه به صورت تصادفی از میان داوطلبان کنکور از میان رشته های ادبیات انگلیسی، آموزش، و مترجمی انتخاب شده اند. در این مطالعه برای بررسی عملکرد افتراقی سوالات در میان داوطلبان مرد و زن از مدل راش استفاده شده است. نتایج نشان داد سوال های کنکور در میان شرکت کننده های زن و مرد، عملکرد افتراقی داشته اند. همچنین آزمون تک بعدی بودن نشان داد که برای پاسخ صحیح دادن به سوالات نیاز به دانش، توانایی و مهارت هایی غیر از دانش زبانی می باشد، بنابراین سوالات آزمون تک بعدی نمی باشد. می توان نتیجه گرفت که نمره های آزمون کنکور دارای واریانس نامربوط به ساختار می باشد و عادلانه بودن کلی آزمون مورد تایید قرار نگرفت. می توان اذعان کرد که نتایج این تحقیق برای طراحی سوالات کنکور، سرپرستان و مسئولین این امر و همچنین معلمان و دانش آموزان حائز اهمیت است.	<p>تاریخ ارسال: ۹۷/۱۲/۲۶</p> <p>تاریخ پذیرش: ۹۸/۰۲/۱۵</p> <p>تاریخ انتشار: ۹۸/۱۲/۲۵</p> <p>نوع مقاله: علمی پژوهشی</p> <p><b>کلید واژگان:</b></p> <p>آزمون های سرنوشت ساز در مقیاس بالا، استقلال موضعی، بعدیت، عملکرد افتراقی آیتم ها، عادلانه بودن، مدل راش</p>

کلیه حقوق محفوظ است ۱۳۹۹

شناسه دیجیتال 10.22059/jflr.2019.278170.611:doi

بردبار، سوده. (۱۳۹۹). بررسی تفاوت عملکرد آیتم ها در آزمون ورودی دانشگاه با استفاده از آنالیز مدل راش. پژوهشهای زبانشناختی در زبانهای خارجی، ۱۰(۱)، ۳۷-۵۵. doi: 10.22059/jflr.2019.278170.611



## JOURNAL OF FOREIGN LANGUAGE RESEARCH

Print ISSN: 2588-4123

Online ISSN: 2588-7521

Website: [www.jflr.ut.ac.ir](http://www.jflr.ut.ac.ir)



### Differential Item Functioning of University Entrance Exam: Using Rasch Analysis

**Soodeh Bordbar**

PhD of TEFL, Iran  
University of Medical  
Science, Tehran, Iran



*Soodeh Bordbar received her M.A and PhD degree in Teaching English as a Foreign Language (TEFL) at University of Tehran, 2009 and 2017*

*Email: [ut.sbordbar@yahoo.com](mailto:ut.sbordbar@yahoo.com)*

#### ARTICLE INFO

##### Article history:

Received 17 March 2019

Accepted 5 May 2019

Available online April 2020

Article Type Research article

##### Keywords:

*Differential Item Functioning, Dimensionality, Fairness, High stakes Test, Local Independence, Rasch Model.*

#### ABSTRACT

The present study aims at investigating the presence of Differential Item Functioning (DIF) in terms of gender in a high stakes language proficiency test, the National University Entrance Exam for Foreign Languages (NUEEFL). The participants (N = 5000) of this study have been selected randomly from a pool of examinees who had taken the NUEEFL as a university entrance requirement for English language studies (English literature, Teaching, and Translation). The results revealed that among 95 items, 40 items exhibit DIF between male and female. Our investigation revealed that the test is not unidimensional and a correct answer requires other knowledge, ability, and skill than the ones that the items aim to measure. It is concluded that the NUEEFL test's scores are not free of construct-irrelevant variance and the overall fairness of the test is under question. In addition, the current research provides several important implications for test designers, stake-holders, administrators, as well as teachers and students.

DOI:10.22059/jflr.2019.278170.611

© 2020 All rights reserved.

Bordbar, Suede. (2020). Investigating the Difference of Items Performance in University Entrance Examination Using Rash Model Analysis. *Journal of Foreign Language Research*, 10(1), 37-55. doi: 10.22059 / jflr.2019.278170.611

## - مقدمه

درباره انتخاب و پذیرش عادلانه در دانشگاه ها محکوم به شکست خواهد بود.

هدف مطالعه حاضر، به طور کلی، بررسی اعتبار آزمون ورودی دانشگاه و به طور خاص، بررسی نقش جنسیت به عنوان منشا سوگیری می باشد. بدون در نظر گرفتن محتوای مباحث درباره مسئله جنسیت، بنظر می رسد هیچ گونه مدرکی دال بر تاثیرات جنسیت در عملکرد NUEEFL وجود ندارد. عامل جنسیت در ارزیابی آزمون مورد توجه ویژه می باشد. در شرایطی که متغیر جنسیت تاثیر زیادی در آزمون نشان دهد، مورد سو گیرانه بشمار می آید و اعتبار آزمون را زیر سوال می برد. زیرا که جنسیت بخشی از ساختار ارزیابی شده توسط آزمون نمی باشد و هر گونه تاثیر قابل توجهی توسط جنسیت عامل واریانس نامربوط به ساختار (Construct- irrelevant Variance) می باشد. با نظر به موارد مذکور و به عنوان بخشی از یک فرایند استاندارد، آنالیز عملکرد افتراقی آیتم ها (DIF) (Differential Item Functioning) در پرسشهای آزمون کنکور به عنوان فاکتور اصلی در ارزیابی عادلانه بودن و اعتبار آزمون های انجام می شود.

## ۲- پیشینه پژوهش

## ۲-۱- آنالیز عملکرد افتراقی آیتم ها (DIF)

آزمون سازان جهت حصول اطمینان از این که پرسش های آزمون برای تمامی امتحان دهندگان مناسب و منصفانه است، کنترل کیفی مداوم یا استفاده از رویکردهای آماری مختلفی را اتخاذ کرده اند (کامیلی و پنفیلد، ۱۹۹۷؛ هالند و وینر ۲۰۱۲؛ رمزی، ۱۹۹۳) (Camilli & Penfield, 1997; Holland & Wainer, 1993; Ramsey, 2012). هدف از رویکرد آماری، شناسایی سوالات آزمون با ویژگی های آماری در گروه های معین آزمودنی است. این مسئله به آنالیز تفاوت عملکرد آیتم ها (DIF) اشاره دارد که " چنین پرسش هایی در گروه های مختلف عملکرد متفاوتی

استفاده از یک آزمون و پیامدهای حاصل از یک تست بحث های زیادی در آزمون سازی زبان و سنجش های آموزشی مطرح می کند. آزمون ابزارهای سنجش از جمله سنجش روایی و پایایی آن ها، دارای اهمیت ویژه ای می باشد. همانطور که شاهد هستیم، تلاش برای ایجاد مجموعه ای از اصول و روش ها، جهت پیوند بین نمره های آزمون و تفاسیر مبتنی بر نمره، برای استفاده از امتحان ها و پیامدهای استفاده از یک آزمون وجود دارد. بنابراین استفاده از یک آزمون بخصوص آزمون های توانایی سنجی زبان انگلیسی بالاخص در بعد کلان در مرکز توجه ارزیابی قرار دارد.

آزمون ورودی دانشگاه دولتی زبان های خارجی (NUEEFL National University Entrance Exam for Foreign Languages) سالانه برای تعداد زیادی از متقاضیان در سراسر ایران برگزار میشود که پیامدهای عدم موفقیت در این آزمون بسیار جدی می باشد. عدم قبولی می تواند تبعات روانی همچون ایجاد افسردگی و سرخوردگی بر اثر ناکامی در قبولی در رشته های دانشگاهی به همراه داشته باشد. همچنین منجر به صرف یک یا دو سال جهت آماده شدن برای کسب رتبه های قبولی یا برتر آزمون گردد. به علاوه در خصوص عدم قبولی آقایان در دانشگاه، دو سال خدمت سربازی را به همراه خواهد داشت.

با توجه به پیامدهای مذکور، بررسی ابزارهای ارزیابی و نتایج حاصل از بررسی اعتبارسنجی و انسجام آزمون (پایایی و روایی آزمون) امری ضروری می باشد (پائنه، ۲۰۱۱) (Pae, 2011). بالاخص ضرورت این امر در سنجش توانایی زبان در آزمون های سرنوشت ساز مانند آزمون ورودی دانشگاه مشاهده می شود. با این وجود، علی رغم ماهیت حساس و بحث داغ اعتبارسنجی آزمون های سرنوشت ساز در مقیاس بالای شرکت کننده، شواهد محدود تجربی ای برای اثبات اعتبار و عادلانه بودن آزمون ورودی دانشگاه وجود دارد. در نبود شواهدی محکم و مستدل، هر گونه صحبت

احتمال متفاوتند. در این صورت می‌توان نتیجه‌گیری کرد که وجود DIF بین گروه‌ها نشان می‌دهد که گروه‌ها نمی‌توانند بطور معناداری در یک پرسش مقایسه شوند (فر و باکارک، ۲۰۰۷)

آنالیز DIF به منظور تعیین اینکه آیا یک آیت یا یک سوال در آزمون، عملکرد یکسانی برای دو یا چند گروه از آزمون‌دهندگان دارد استفاده می‌شود؛ که "معمولا با پیش‌زمینه نژادی، مذهبی، جنسیت، سن، تجربه و یا شرایط معلولیت تعریف میشوند" (شونمان و بلیستین، ۱۹۸۹، ص. ۲۵۵-۲۵۶) (Scheuneman & Bleistein, 1989, pp. 255-6). تعداد چشمگیری از مطالعات، شگردهای تشخیص و یافتن DIF را طبقه‌بندی کرده‌اند. و تاکنون بسیاری از شگردهای تحلیل DIF ارائه شده‌اند. مکنومارا و روور (۲۰۰۶، ص. ۹۳) (McNamara & Roever, 2006, p. 93) روش‌های تشخیص و یافتن DIF را به چهار گروه کلی و ۹۳ جامع طبقه‌بندی کرده‌اند: (۱) تحلیل بر پایه دشواری سوال/آیتم؛ این رویکردها تخمین‌های سختی سوال را مقایسه می‌کنند. (۲) رویکردهای ناپارامتری: در این روش‌ها از جدول‌های وابستگی احتمالی (Contingency tables)، مجذور خی (کای اسکور) و نسبت یا ضریب احتمال استفاده می‌شود. (۳) رویکردهایی بر پایه نظریه- پاسخ- پرسش (Item-response theory-based approaches (IRT): که شامل آنالیزهای ۱، ۲ و ۳ پارامتری IRT میشوند. (۴) سایر رویکردها: که شامل رگرسیون لاجیستیک می‌شود که یک روش مدل مقایسه‌ای را بکار می‌برد. همچنین نظریه تعمیم‌پذیری و سنجش چند وجهی، که در مطالعات کلاسیک DIF کمتر مورد استفاده قرار می‌گیرند. بسیاری از شگردهایی که قابل اجرا نیستند، موجود می‌باشند، اگرچه تنها تعداد محدودی از آن‌ها اخیرا مورد استفاده قرار گرفته است. بخش ذیل مدل‌هایی که بر پایه نظریه IRT، بخصوص مدل راش (Rasch Model) را به‌عنوان روش قابل اجرا و مرتبط برای تحقیق حاضر در نظر گرفته و مورد بررسی قرار می‌دهد.

دارند که اشاره به شاخص بالقوه سوگیری و تبعیض در سوالات دارد" (سیرسی و رایوس، ۲۰۱۳ و ص. ۱۷۰) (Sireci & Rios, 2013, p. 170).

بر طبق نظریه گرانیپایه و کونان (۲۰۰۷) (Geranpayeh & Kunnan, 2007) صرف نظر از بررسی مسئله عادلانه بودن در چرخه طراحی-ایجاد-اجرا-نمره دهی، هنوز مشکلات بسیاری در این پروسه یافت می‌شود. گرانیپایه و کونان (۲۰۰۷) به بیان رویکردی مناسب برای حل این مشکلات می‌پردازند. این رویکرد پیشنهاد ایجاد یک گروه آزمایشی به منظور بررسی نمره‌های آزمون می‌دهد. چنانچه آزمونی برگزار شده باشد، لازم است نمونه‌ای بزرگ جهت بررسی نمره‌های آزمون و عملکرد آیت‌ها مورد استفاده قرار گیرد. در صورتی که مشخص شود آنها به صورت متفاوت عمل کرده‌اند، منبع این تفاوت به‌عنوان عملکرد افتراقی آن سوال شناخته می‌شوند.

برای تبیین تعریف واژه DIF، ویبرگ (۲۰۰۷) (Wiberg, 2007) اشاره می‌کند که شناسایی آیتم‌های مشکل ساز از طریق تجزیه و تحلیل سوال‌ها، نقش مهمی در یک آزمون ایفا می‌کند. همچنین ویبرگ اذعان داشت "تجزیه و تحلیل سوال‌ها شامل استفاده از شگردهای آماری به منظور بررسی عملکرد آزمون‌دهندگان در آیت‌ها میشود" (ویبرگ، ۲۰۰۷، ص. ۱) که یکی از ضروری‌ترین بخش‌ها در تجزیه و تحلیل سوال‌ها، شناسایی عملکردهای متفاوت آیت‌ها میباشد. شگرد DIF هنوز یک روش مفید برای شناسایی پرسش‌های مشکل ساز تلقی میشود.

عملکرد افتراقی آیت‌ها زمانی روی می‌دهد که "ویژگی‌های یک سوال در یک گروه، متفاوت با ویژگی‌های پرسش در گروه دیگر باشد" (فر و باکارک، ۲۰۰۷، ص. ۳۳۱) (Furr & Bacharach, 2007, p. 331). به منظور تبیین موضوع، فر و باراک مسئله را با ذکر یک مثال مشخص کردند. آن‌ها معتقد هستند که DIF زمانی بوجود می‌آید که یک پرسش خاص دارای سطوح متفاوتی از سختی برای مردان و زنان باشد. به بیان دیگر وقوع عملکرد متفاوت آیت‌ها بدین معناست که یک زن و یا یک مرد با سطح توانایی و خصوصیت‌های یکسان، در پاسخگویی صحیح به سوالات دارای

فاکتورهای نامربوط به ساختار، همچون جنسیت، قومیت و پیشینه تحصیلی را از طریق محاسبه تفاوت عملکرد آیتم ها فراهم می‌کند.

در ادامه به فرضیات مدل راش که شامل تک بعدی بودن و استقلال موضعی (Unidimensionality and Local Independence) می‌باشد خواهیم پرداخت. همانطور که از اسم این آزمون نمایان است، آزمون تک بعدی شامل پرسش‌هایی می‌شود که تنها یک بعد دارند. در این خصوص، دی مارس (۲۰۱۰، ص. ۳۸) اذعان داشت "هنگامی که فقط یک نمره برای یک آزمون گزارش می‌شود، تلویحا این فرضیه وجود دارد که پرسش‌ها در یک ساختار اولیه مشترک سهیم هستند". فرضیه تک بعدی بودن مستلزم این است که "آیتم‌ها هم‌آهنگ و متحد عمل کنند و تلم و واریانس‌های غیرتصادفی در داده‌ها می‌توانند از طریق توانایی و دشواری سوال در نظر گرفته شوند" (ویل، ۲۰۱۳، ص. ۵۶) (Wale, 2013, p. 56). بطور کلی تک بعدی بودن حاکی از این است که آیا آیتم‌ها ایجاد یک ویژگی نهفته را می‌سازند یا خیر.

در تک بعدی بودن باید با هشجاری و احتیاط بیشتری برخورد کرد. به‌گونه‌ای که پاسخ‌ها به پرسش‌های آزمون می‌توانند از نظر ریاضی تک بعدی باشند، در حالی که آیتم آزمون آنچه که معلم و روانشناس به‌عنوان دو ساختار مجزا در نظر دارند را می‌سنجد. به‌عنوان مثال پرسش‌های آزمون ممکن است سرعت و دانش پاسخگویی را اندازه‌گیری کنند.

فرضیه دیگر مدل راش استقلال موضعی است. ویل (۲۰۱۳) استقلال موضعی را به‌این‌گونه تبیین می‌کند که احتمال پاسخ صحیح دادن آزمون شونده به یک سوال خاص و مشخص به پاسخ‌های قبل او و یا به پاسخ‌های افراد دیگر به‌همان سوال وابسته و مرتبط نمی‌باشد.

تک بعدی بودن می‌تواند از طریق آمار برازش مدل بررسی شود. به علاوه، تک بعدی بودن و استقلال موضعی با استفاده از آمار مدل برازش تخمین زده می‌شوند. چنانچه شخص یا آیتمی برازش متناسب نداشت، چنین تفسیر می‌شود که به‌چه‌میزانی یک شخص

IRT که در امتداد تعمیم نظریه کلاسیک آزمون سنجی با ریشه‌های ریاضی است، عمیقا در روانشناسی ریشه دوانده است و اصول ریاضی آن در روانسنجی گنجانده شده است (استینی و نرینگ، ۲۰۰۶) (Ostini & Nering, 2006). برخی موضوعات بحث برانگیز در تعریف مفهوم سنجش در رشته‌های علوم انسانی و روانشناسی وجود دارد. تعدادی از صاحب نظران بر این باورند که مدل راش از نظر ریاضی با مدل یک- پارامتری لاجیستیک IRT برابر است، اما باید توجه داشت که هر کدام به‌صورت جداگانه ایجاد و گسترش پیدا کردند (دی مارس، ۲۰۱۰) (DeMars, 2010). در این ارتباط بحث و اختلاف نظرهایی پیرامون مدل راش وجود دارد. همان‌طور که بعضی از متخصصان این علم براین باورند، مدل راش و مدل‌های IRT از نظر ساختاری متفاوت هستند و به‌صورت کاملا متمایزی عمل می‌کنند. در توضیح این ادعا باید گفت که مدل‌های IRT به‌منظور توصیف و برازش داده‌ها بکار می‌روند و زمانی که داده‌ها برازش ضعیفی داشته باشند، مدل مذکور به‌نفع مدل دیگر تعدیل یا حذف می‌شود. در مقابل مدل راش، بیشتر به‌صورت تجویزی اعمال می‌شود. بدین صورت که داده‌ها باید با مدل برازش، مناسبت داشته باشند و هنگامی که برازش ایجاد نشود آیتم‌هایی که با مدل عدم تناسب دارند تا زمانی که برازش رضایت بخش بدست نیاید حذف می‌شوند (زند شولتن، ۲۰۱۱، ص. ۳۹) (Zand Scholten, 2011, p. 39).

مدل راش شامل یک سنجش مدل-محور است، که در آن ارزیابی سطح توانایی وابسته به پاسخ‌های افراد و ویژگی‌های آیتم‌هایی که در آن آزمون تدوین شده اند، می‌باشد (امبرتسون و ریز، ۲۰۰۰، ص. ۱۳) (Embretson & Reise, 2000, p. 13). علاوه براین، پرسش‌های آزمون نباید برای هیچ یک از زیر مجموعه‌های شرکت کنندگان متفاوت عمل کنند. چنانچه یک پرسش برای یک‌گروه عملکرد متفاوتی داشته باشد، اعتبار ارزیابی برای یک ساختار خاص کاهش می‌یابد و این مسئله به‌عنوان تهدیدی برای عادلانه و بی‌غرض بودن آزمون تلقی می‌شود. رویکرد مدل راش، امکان بررسی پرسش‌های سوگیر نسبت به زیرگروه‌های متفاوت و بررسی

مورد تحلیل آماری قرار گرفت و با در نظر گرفتن هدف مطالعه تنها زیر بخش‌های زبان تخصصی گزارش خواهد شد.

این آزمون هر سال برای بیش از ۱۰۰۰۰۰ متقاضی دانشگاه که به دنبال تحصیل در یکی از رشته‌های زبان‌های خارجی هستند برگزار می‌شود. تمامی سوالات چند گزینه ای میباشند و به صورت دو قطبی نمره دهی میشوند. مدت زمان آزمون ۱۰۵ دقیقه میباشد. بطور کلی به عنوان یک قاعده در آزمون NUEEFL؛ حدس زدن مجاز نمی‌باشد و نمره منفی برای پاسخ‌های اشتباه در نظر گرفته می‌شود. بدین معنا که ۳ آیتم اشتباه یک پاسخ درست را حذف می‌کند.

در بخش سنجش‌های آماری، نسخه نرم‌افزار وینستپ ورژن ۳,۹۲,۱ (Winsteps software Version 3.92.1) برای تحلیل داده‌ها بکار گرفته شده است. این نرم افزار سنجش راش را از مجموعه داده‌های ساده مانند افراد و آیتم‌ها می‌سازد و مدل راش دو قطبی را بکار می‌برد. همچنین پایایی آزمون- پیرسون و پایایی پرسش این آزمون عالی گزارش شده است (پیرسون  $r = 0,93$  و  $r = 1$ ).

### ۳-۳- فرایند تحقیق

هرساله NUEEFL برای شمار بسیاری از شرکت‌کننده‌ها در سراسر ایران برگزار می‌شود. مطالعه فعلی بر یک جنبه اعتبار آزمون که از طریق انجام مدل راش ارزیابی می‌شود، متمرکز شده است. برای انجام آنالیز DIF و اعمال مدل راش می‌باید درستی فرض‌های آماری و ریاضی که پیش‌تر بدان اشاره شد، مورد بررسی قرار گیرد.

### ۳-۴- آنالیز داده‌ها

ویژگی‌های روانسنجی آیتم‌ها با استفاده از نرم افزار وینستپ (لیناکر، ۲۰۱۶ ب) (Linacre, 2016b) تخمین زده شده است. مجموعه داده‌ها دو قطبی می‌باشد. در این نرم افزار برای برآورد

و یا سوال مورد نظر، آن چنان که مدل راش پیش‌بینی کرده است عمل نمی‌کند.

### ۲-۳- پرسش‌های اصلی پژوهش

بمنظور بررسی ویژگی‌های روان‌سنجی آزمون در مقیاس و ریسک بالا، مطالعه حاضر به پرسش‌های زیر می‌پردازد:

۱. تا چه اندازه‌ای پاسخ‌ها به سوال آزمون NUEEFL، یک ساختار تک بعدی را بر طبق مدل ارزیابی راش شکل می‌دهد؟

۲. آیا جنسیت شرکت‌کنندگان یک منبع DIF در پرسش‌های NUEEFL می‌باشد؟

### ۳- روش پژوهش

#### ۳-۱- شرکت کنندگان

از میان ۲۰۰۰۰ شرکت‌کننده در آزمون NUEEFL تعداد ۵۰۰۰ نفر انتخاب شدند. آزمودنی‌ها به صورت تصادفی از دو گروه با جنسیت متفاوت برگزیده شدند. گروه زنان ۳۳۳۵ و مردان ۱۶۶۵ نفر از کل جمعیت شرکت‌کنندگان را تشکیل می‌دهد. در تحقیق حاضر، پیشینه تحصیلی و سن شرکت‌کنندگان در نظر گرفته نشده است.

#### ۳-۲- ابزار

وسیله‌ی مورد استفاده در تحقیق حاضر، بخشی از امتحان ورودی دانشگاه دولتی برای زبان‌های خارجه (NUEEFL) است. در قسمت زبان انگلیسی این آزمون ۹۵ پرسش چهارگزینه‌ای است. از میان ۹۵ آیتم، ۲۵ پرسش زبان عمومی و ۷۰ پرسش تخصصی است که از شش زیر مجموعه تشکیل می‌شود. زیر مجموعه‌های آزمون تخصصی شامل ۱- دستور زبان انگلیسی (۱۰ سوال)، ۲- کلمات (۱۵ سوال)، ۳- ساختار جمله (۵ سوال)، ۴- عملکردهای زبانی (۱۰ سوال)، ۵- آزمون کلوز (۱۵ سوال) و ۶- درک مطلب (۱۵ سوال). در تحقیق حاضر هر دو بخش عمومی و تخصصی

پارامترها از روش بیشینه درست نمایی توام (JMLe) (Joint Maximum Likelihood Estimation) استفاده می‌گردد.

در فرمول JMLe، تخمین پارامتر راش زمانی صورت می‌گیرد که نمره خام مشاهده شده برای پارامتر با نمره خام مورد انتظار مطابقت داشته باشد.

برای تخمین برازش سوال‌ها با مدل (Data-model fit estimation) از طریق به‌کارگیری مقادیر مربع میانگین infit و outfit برای شناسایی آیت‌های misfit (عدم برازش) و goodfit (برازش خوب) استفاده می‌شود. زمانی که گفته می‌شود یک فرد یا یک آیت ممکن است عدم برازش (misfitting) داشته باشد، بدان معناست که شخص یا آیت مورد نظر آن چنان که مدل راش پیش بینی کرده است، عمل نمی‌کند (بونه، استیور و ییل، ۲۰۱۴) (Boone, Staver, & Yale, 2014).

نرم‌افزار وینستپ برای بررسی برازش سوال‌ها با مدل، دو نوع آماره MnSq و Zstd را ارائه می‌دهد. دامنه آماره MnSq از ۰ تا بی‌نهایت و مقدار مورد انتظار آن ۱ است. مقادیر بالاتر از ۱ بیانگر انحراف از تک بعدی بودن، و مقادیر پایین‌تر از ۱ حاکی از بیش‌برازش الگوهای پاسخ با مدل است که به معنای وجود وابستگی در میان پاسخ‌ها یا پرسش‌هاست. آماره Zstd که از تبدیل آماره MnSq به یک آماره t به دست می‌آید، در نمونه‌های بزرگ دارای توزیع Z است. در واقع این آماره، معناداری آماری انحراف داده‌ها از مدل راش را مورد بررسی قرار می‌دهد. لازم به یادآوری است که این آماره به حجم نمونه بسیار حساس است. برای نمونه‌هایی با حجم کوچکتر از ۹۰ نفر هر نوع داده‌ای با مدل برازش می‌یابد، در حالی که برای نمونه‌های بزرگتر از ۹۰۰ نفر هیچ نوع داده‌ای با مدل برازش نمی‌یابد. بنابر این با توجه به حجم بالای نمونه مورد استفاده در این پژوهش و همچنین با در نظر گرفتن توصیه لینکر (۲۰۱۲) (Linacre, 2012)، جهت بررسی برازش سوال با مدل از آماره MnSq که به دو شکل outfit و infit برآورد و گزارش می‌شود استفاده گردید.

از آن جایی که لینکر (۲۰۱۲) آماره outfit MnSq را بر آماره infit MnSq ترجیح می‌دهد. بنابراین در پژوهش حاضر برای بررسی برازش سوال از آماره outfit MnSq استفاده شد. دامنه قابل قبول برای این آماره از ۰٫۷ تا ۱٫۳ می‌باشد. در متد اندازه‌گیری راش برای بررسی تک بعدی بودن سؤال‌ها از روش‌های متعدد، از جمله آماره‌های برازش که در بالا ذکر شد، استفاده می‌شود. با این حال یافته‌های پژوهشی نشان می‌دهد که این آماره‌ها از حساسیت لازم برای شناسایی چند بعدی بودن برخوردار نیستند، بنابراین منطقی است که در کنار آماره‌های برازش از تحلیل مؤلفه‌های اصلی روی داده‌های خام یا پسمانده‌ها (Principal Component Analysis (PCA)) نیز استفاده کرد. در پژوهش حاضر، برای بررسی تک بعدی بودن سؤالاها از روش تحلیل مؤلفه‌های اصلی روی پسمانده‌ها، و آزمون t استفاده گردید. در روش PCA برای تعیین تک بعدی بودن سوال‌ها از ملاک‌های پیشنهادی لینکر (۱۹۹۱-۲۰۰۶) (Linacre, 1991-2006, p. 272) استفاده شد که شامل موارد زیر می‌باشد:

۱. عامل راش یعنی صفت مورد اندازه‌گیری f حداقل ۰٫۶۰٪ از واریانس را تبیین نماید.
۲. مقدار ویژه اولین مقابله (Eigen Value) (یعنی اولین مولفه پس مانده) کمتر از ۳ باشد و کمتر از ۵٪ از واریانس را تبیین نماید.

#### ۴- نتایج

##### ۴-۱- پارامتر دشواری و برازش سؤال‌ها

همان گونه که توضیح داده شد نرم افزار وینستپ برای بررسی برازش سوال‌ها با مدل راش، دو نوع آماره MnSq و Zstd را ارائه می‌دهد. دامنه آماره MnSq از ۰ تا بی‌نهایت و مقدار مورد انتظار آن ۱ است. مقادیر بالاتر از ۱ بیانگر انحراف از تک‌بعدی بودن، و مقادیر پایین‌تر از ۱ حاکی از بیش‌برازش الگوهای پاسخ با مدل است که به معنای وجود وابستگی در میان پاسخ‌ها یا سوالات است. با توجه به توضیحات در قست قبل، در این تحقیق برای

**MnSq** استفاده گردیده است. دامنه مقادیر این آماره از ۰,۵۷ تا ۳,۳ است که حاکی از این است که برخی از سئوال‌ها با مدل راش برازش ندارند. بررسی آماره **Outfit MnSq** نشان می‌دهد که تعداد ۲۶ سئوال (یعنی ۲۷ درصد از مجموع سئوال‌ها) فاقد برازش با مدل راش‌اند. به‌علت محدودیت فضا، محقق در جدول زیر به‌گزارش پارامتر دشواری و برازش سئوال‌هایی که دارای عدم برازش (**Misfitting Items**) با مدل راش دارند اکتفا کرده است.

بررسی برازش سئوال‌ها از آماره **outfit MnSq** استفاده گردید. دامنه قابل قبول برای این آماره از ۰,۷ تا ۱,۳ می‌باشد.

برآوردهای مربوط به دشواری سئوال‌ها همراه با خطای استاندارد مدل، پارامتر تشخیص و همچنین مقادیر آماره‌های **Outfit MnSq** و **Infit MnSq** در جدول ۱ آمده است. دامنه پارامتر دشواری از ۲,۹۹- تا ۲,۴۵ با میانگین ۰ و انحراف استاندارد ۱,۲۱ می‌باشد.

اگرچه در جدول زیر مقادیر آماره **infit MnSq** نیز ارائه شده است، لکن در ارزیابی برازش سئوال‌ها صرفاً از آماره **Outfit**

جدول ۱. پارامتر دشواری و برازش سئوال‌ها با عدم برازش برای آزمون NUEEFL

Item	Entry Number	Total Score	Measure	Model S.E.	Infit MNSQ	infit ZSTD	Outfit MNSQ	Outfit ZSTD
Q155	80	158	2.45	0.08	1.07	1	1.82	4.8
Q126	51	191	2.23	0.08	1.08	1.2	2.05	6.3
Q137	62	265	1.84	0.07	1.04	0.8	1.32	2.6
Q105	30	285	1.76	0.07	1.19	3.7	3.01	9.9
Q118	43	314	1.64	0.06	1.18	3.6	1.98	7.3
Q166	91	330	1.57	0.06	1.05	1.2	1.49	4.2
Q101	26	335	1.56	0.06	1.26	5.3	3.3	9.9
Q103	28	363	1.46	0.06	1.28	6.1	2.99	9.9
Q158	83	367	1.44	0.06	1.14	3.1	1.46	4.2
Q111	36	392	1.36	0.06	1.12	2.9	1.87	7.4
Q115	40	411	1.3	0.06	1.14	3.5	2.02	8.6
Q133	58	411	1.3	0.06	1.1	2.5	1.77	6.8
Q121	46	459	1.15	0.05	1.38	9.2	2.43	9.9
Q167	92	462	1.14	0.05	0.83	-4.9	0.64	-4.7
Q109	34	463	1.14	0.05	1.24	6	2.03	9.2
Q128	53	496	1.05	0.05	1.2	5.3	1.44	4.7
Q122	47	600	0.79	0.05	1.13	4.2	1.46	5.4
Q156	81	732	0.5	0.04	1.2	6.9	1.41	5.4
Q99	24	821	0.33	0.04	0.83	-7.1	0.69	-5.7



Item	Entry Number	Total Score	Measure	Model S.E.	Infit MNSQ	infit ZSTD	Outfit MNSQ	Outfit ZSTD
Q84	9	860	0.26	0.04	1.28	9.9	1.6	8.5
Q153	78	874	0.24	0.04	0.78	-9.4	0.59	-8.1
Q149	74	1046	-0.05	0.04	0.75	-9.9	0.57	-9.9
Q108	33	1079	-0.1	0.04	1.17	7.6	1.33	6.1
Q160	85	1151	-0.21	0.04	0.8	-9.9	0.67	-8.1
Q91	16	2076	-1.37	0.03	0.76	-9.9	0.67	-9.9
Q79	4	2513	-1.85	0.03	1.2	9.9	1.36	9.9
Mean		1170.9	0	0.04	1	-0.7	1.14	0.1
P.SD		790.5	1.21	0.01	0.13	5.4	0.5	5.3

\* نکته: محاسبات از تعداد کل ۵۰۰۰ نفر می‌باشد. که بسیار پایین تر از ملاک پیشنهادی لی ناگر قرار دارد. همچنین مقدار ویژه اولین تا پنجمین عامل باقیمانده، به ترتیب ۳,۴ - ۲,۵ - ۲,۲ - ۱,۹ - ۱,۷ بوده که بالاتر از مقدار پیشنهادی لینگر قرار دارد.

#### ۲-۴- بعدیت آزمون و استقلال موضعی

برای آزمون تک بعدی بودن سوال ها علاوه بر آماره‌های برازش سوال، از تحلیل مولفه‌های اصلی روی پس مانده‌های استاندارد شده و همچنین آزمون t استفاده شد. نتایج حاصل از تحلیلی مولفه‌های اصلی بر روی پس مانده ها نشان داد که عامل راش ( یعنی پارامتر مورد اندازه گیری آزمون) ۳۴,۸ درصد از واریانس را تبیین می‌کند

#### جدول ۲. آنالیز PCA

مورد انتظار	مشاهده شده	مقدار ویژه	واریانس در واحدهای مقدار ویژه
٪۳۳,۷	٪۳۴,۸	۵۰,۷۰۶۸	واریانس خام تبیین شده توسط اندازه ها
٪۱۲,۳	٪۱۲,۷	۱۸,۵۴۵۴	واریانس خام تبیین شده توسط افراد
٪۲۱,۴	٪۲۲,۱	۳۲,۱۶۱۴	واریانس خام تبیین شده در آیتم ها
٪۶۶,۳	٪۶۵,۲	۹۵,۰۰۰۰	واریانس تبیین نشده خام
٪۱۰۰,۰	٪۱۰۰,۰	۱۴۵,۷۰۶۸	واریانس خام کل در مشاهدات

جدول ۳. مقدار ویژه اولین تا پنجمین عامل باقیمانده

مورد انتظار	مشاهده شده	مقدار ویژه	تقابل در واحدهای مقدار ویژه
٪۳,۶	٪۲,۴	۳,۴۶۶۲	واریانس تبیین نشده از اولین تقابل
٪۲,۷	٪۱,۷	۲,۵۳۹۷	واریانس تبیین نشده از دومین تقابل
٪۲,۴	٪۱,۵	۲,۲۵۶۰	واریانس تبیین نشده از سومین تقابل
٪۲,۰	٪۱,۳	۱,۹۴۳۸	واریانس تبیین نشده از چهارمین تقابل
٪۱,۸	٪۱,۲	۱,۷۰۷۶	واریانس تبیین نشده از پنجمین تقابل
٪۱۰۰,۰	٪۱۰۰,۰	۹۵,۰۰۰۰	واریانس تبیین نشده خام (کل)

:(Benjamini & Hochberg, 1995)

که در اینجا عامل راش فقط ۳۲,۱ درصد از واریانس را تبیین می‌نماید که بسیار پایین تر از ملاک پیشنهادی لینکر قرار دارد. اما مقدار ویژه اولین مقابله با مقدار ۳,۴ در شرایط اعلام شده صدق می‌کند. به‌طور کلی می‌توان گفت که شرایط لازم برای تک‌بعدی بودن سئوالات برقرار نیست.

برای اطمینان در بررسی بعدیت استقلال موضعی، از آزمون با روش بنجامینی-هاچبرگ استفاده شد. ترتیب و روش آنالیز به‌اختصار به‌شرح زیر است. ابتدا با استفاده از مدل راش پارامتر دشواری برای تمامی آیتم ها محاسبه گردید. بررسی آماره  $Outfit\ MnSq$  نشان داد که مقدار این آماره برای ۲۰ سؤال بیش از ۱,۳۰ است.

۴-۳- بررسی بعدیت و استقلال موضعی آزمون با استفاده از

روش بنجامینی - هاچبرگ

جدول ۴. مرحله اول (A) آزمون تی استیودنت در NUEEFL

Level A				
Item	Entry Number	Total Score	Measure	Outfit MNSQ
Q79	4	2513	-1.85	1.36
Q84	9	860	0.26	1.6
Q101	26	335	1.56	3.3
Q103	28	363	1.46	2.99

Q105	30	285	1.76	3.01
Q108	33	1079	-0.1	1.33
Q109	34	463	1.14	2.03
Q111	36	392	1.36	1.87
Q115	40	411	1.3	2.02
Q118	43	314	1.64	1.98
Q121	46	459	1.15	2.43
Q122	47	600	0.79	1.46
Q126	51	191	2.23	2.05
Q128	53	496	1.05	1.44
Q133	58	411	1.3	1.77
Q137	62	265	1.84	1.32
Q155	80	158	2.45	1.82
Q156	81	732	0.5	1.41
Q158	83	367	1.44	1.46
Q166	91	330	1.57	1.49

با استفاده از تحلیل رانش پارامتر دشواری باردیگر، براساس این آیتم‌ها محاسبه شد، نتایج را در جدول ۵ می‌نگریم.

جدول ۵. مرحله دوم (B) آزمون تی استیوندت در NUEEFL

Level B				
Item	Entry Number	Total Score	Measure	Outfit MNSQ
Q79	1	2513	-2.88	1.23
Q84	2	860	-0.82	1.05
Q101	3	335	0.39	1.13
Q103	4	363	0.3	1.02
Q105	5	285	0.58	0.96
Q108	6	1079	-1.16	0.96

Q109	7	463	0	0.99
Q111	8	392	0.2	0.85
Q115	9	411	0.15	0.95
Q118	10	314	0.47	0.87
Q121	11	459	0.01	1.11
Q122	12	600	-0.33	0.85
Q126	13	191	1.03	0.89
Q128	14	496	-0.08	0.93
Q133	15	411	0.15	0.9
Q137	16	265	0.66	0.8
Q155	17	158	1.24	0.74
Q156	18	732	-0.59	0.98
Q158	19	367	0.28	0.9
Q166	20	330	0.41	0.87

۴-۴- عملکرد افتراقی آیتم‌ها (تغییر ناپذیری یا آنالیز DIF)

گام بعدی در آنالیز داده‌ها، تجزیه و تحلیل DIF است. توانایی و دقت تخمین‌زده شده در مدل رش غیر قابل تغییر فرض می‌شود. هدف روش آماری، تشخیص آیتم‌ها با ویژگی‌های آماری متفاوت در گروه خاصی از افراد مورد مطالعه می‌باشد. تغییرناپذیری سئوال‌ها در این پژوهش، در میان دوگروه زنان و مردان مورد مطالعه قرار گرفت. نتایج حاصل از این تحلیل برای سئوال‌ات در زیر نشان داده شده است. به‌علت محدودیت فضا، فقط آیتم‌هایی که عملکرد افتراقی در بین جنسیت داشتند گزارش شده است.

با استفاده از ثابت تعدیل‌کننده با بدست آوردن میانگین تفاضل بین پارامترهای دشواری مرحله B نسبت به مرحله A برابر ۱,۱۴۲ - محاسبه شد. از سوی دیگر محاسبه پارامتر توانایی افراد بر اساس آیتم‌ها در مراحل A و B انجام شد و در نتیجه از تعداد ۵۰۰۰ آزمون تی استیودنت ۲۶۸۰ آزمون معادل ۵۳,۶ درصد معنی‌دار شده که از میزان ۵ درصد مورد قبول بسیار بالاتر است. بنابراین شرایط لازم برای تک بعدی بودن و استقلال موضعی سئوال‌ات برقرار نیست.

جدول ۶. آنالیز DIF

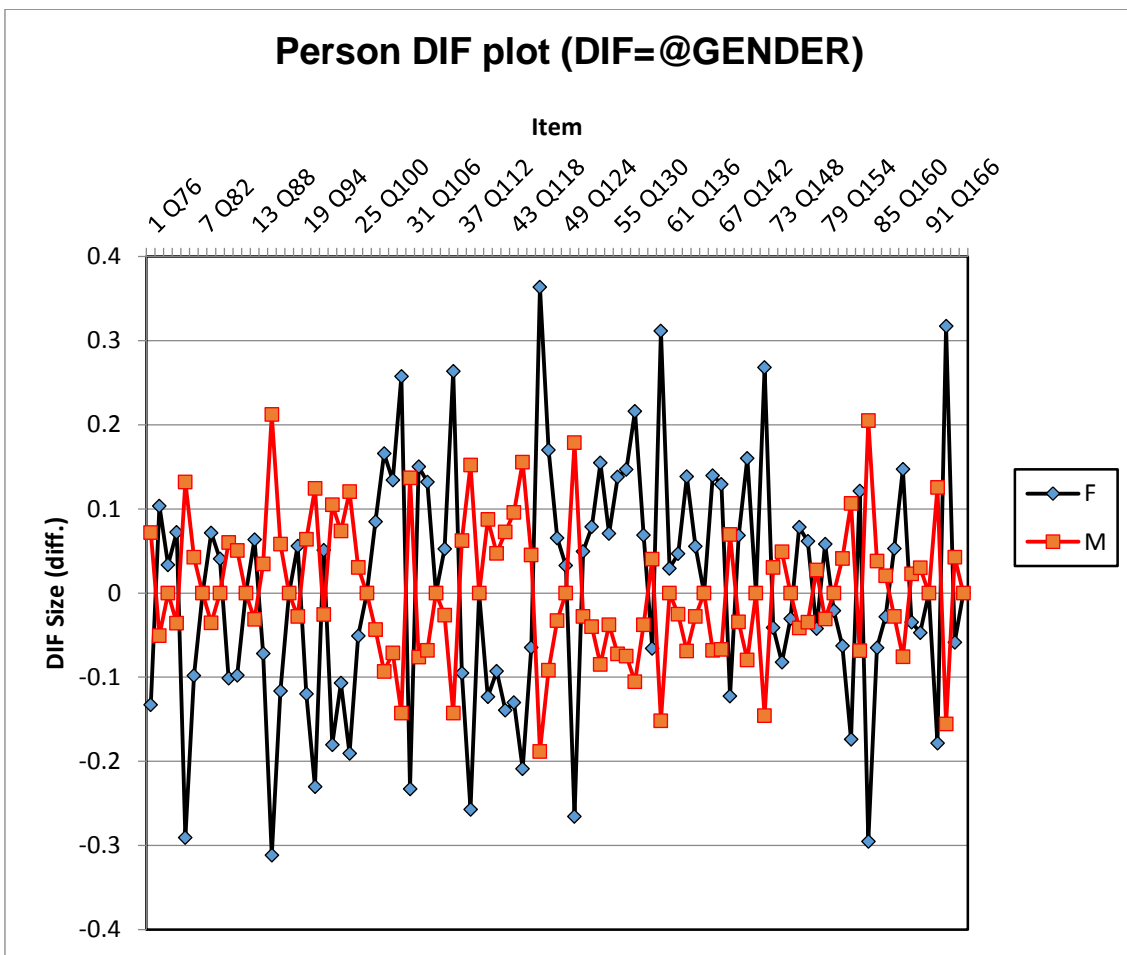
Items	Person Class	DIF Measure	Person Class	DIF Measure	DIF Contrast	Rasch-Welch		
						t	df	Prob
Q76	M	-0.87	F	-0.67	-0.2	-2.75	INF	0.0059
Q77	M	-1.95	F	-2.11	0.15	2.19	INF	0.0289

Items	Person Class	DIF Measure	Person Class	DIF Measure	DIF Contrast	Rasch-Welch		
						t	df	Prob
Q80	M	-2.5	F	-2.08	-0.42	-5.85	INF	0.0000
Q85	M	-0.16	F	0	-0.16	-1.96	INF	0.0498
Q86	M	-1.17	F	-1.02	-0.15	-2.05	INF	0.0400
Q90	M	0.29	F	0.81	-0.52	-5.6	INF	0.0000
Q91	M	-1.48	F	-1.31	-0.17	-2.47	INF	0.0137
Q94	M	-0.99	F	-0.81	-0.18	-2.5	INF	0.0123
Q95	M	-1.08	F	-0.72	-0.35	-4.83	INF	0.0000
Q97	M	-0.41	F	-0.12	-0.29	-3.59	INF	0.0003
Q99	M	0.14	F	0.45	-0.31	-3.52	INF	0.0004
Q103	M	1.62	F	1.36	0.26	2.1	INF	0.0362
Q104	M	0.55	F	0.35	0.21	2.22	INF	0.0263
Q105	M	2.01	F	1.61	0.4	2.89	INF	0.0039
Q106	M	-0.46	F	-0.09	-0.37	-4.66	INF	0.0000
Q107	M	-0.4	F	-0.63	0.23	2.93	INF	0.0034
Q108	M	0.03	F	-0.17	0.2	2.42	INF	0.0157
Q111	M	1.62	F	1.21	0.41	3.35	INF	0.0008
Q113	M	-0.48	F	-0.07	-0.41	-5.15	INF	0.0000
Q117	M	-1.19	F	-0.98	-0.21	-2.93	INF	0.0034
Q119	M	1.31	F	1.67	-0.36	-2.99	INF	0.0028
Q121	M	1.52	F	0.96	0.55	4.77	INF	0.0000
Q122	M	0.96	F	0.7	0.26	2.58	INF	0.0100
Q125	M	0.36	F	0.81	-0.44	-4.71	INF	0.0000
Q128	M	1.2	F	0.96	0.24	2.2	INF	0.0279
Q130	M	0.39	F	0.18	0.21	2.37	INF	0.0179
Q131	M	-0.18	F	-0.41	0.22	2.78	INF	0.0054
Q132	M	-1.38	F	-1.7	0.32	4.55	INF	0.0000
Q135	M	-0.96	F	-1.42	0.46	6.43	INF	0.0000
Q138	M	-0.86	F	-1.07	0.21	2.84	INF	0.0046

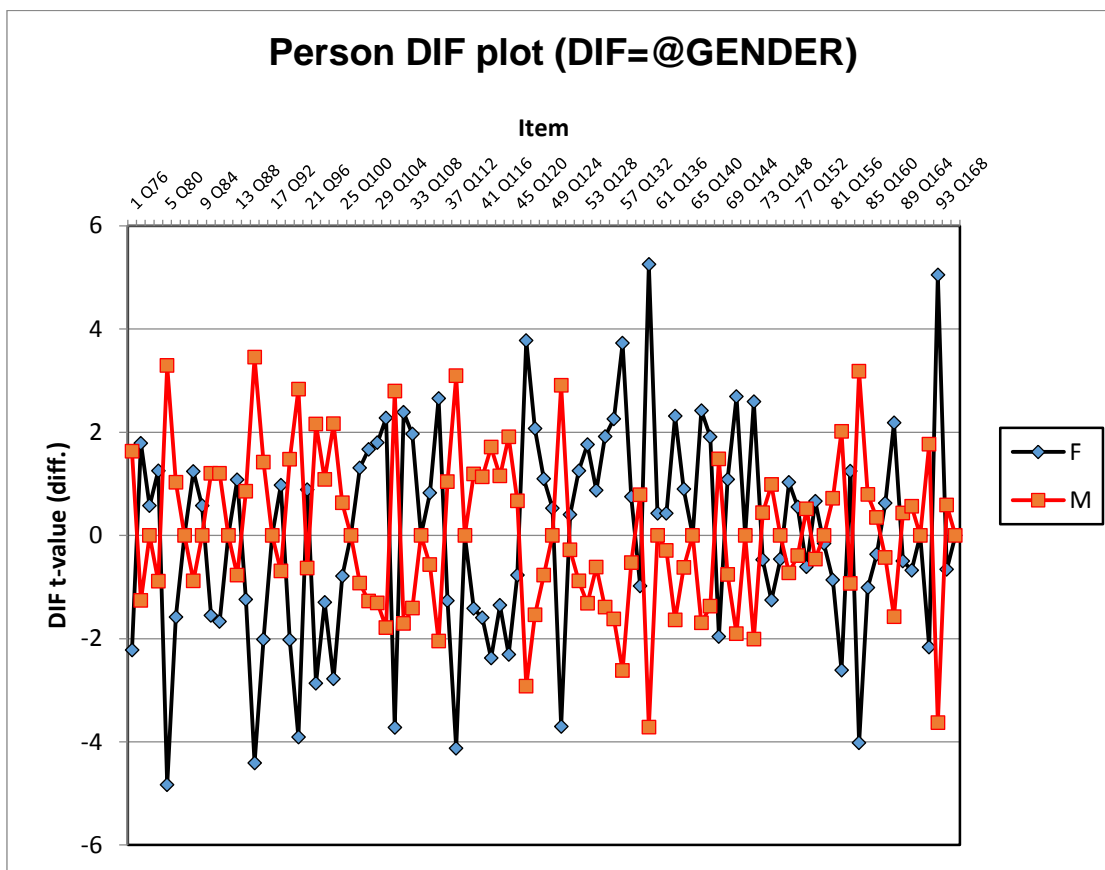
Items	Person Class	DIF Measure	Person Class	DIF Measure	DIF Contrast	Rasch-Welch		
						t	df	Prob
Q141	M	-1.81	F	-2.02	0.21	2.95	INF	0.0032
Q142	M	0.08	F	-0.12	0.2	2.35	INF	0.0187
Q143	M	-0.49	F	-0.3	-0.19	-2.46	INF	0.0140
Q145	M	-0.9	F	-1.14	0.24	3.3	INF	0.0010
Q147	M	1.74	F	1.33	0.41	3.28	INF	0.0011
Q157	M	-0.04	F	0.24	-0.28	-3.3	INF	0.0010
Q159	M	0.47	F	0.97	-0.5	-5.12	INF	0.0000
Q163	M	0.05	F	-0.18	0.22	2.69	INF	0.0072
Q167	M	0.96	F	1.27	-0.3	-2.8	INF	0.0052
Q168	M	-0.42	F	-0.89	0.47	6.22	INF	0.0000

(M = Male (مرد); F = Female (زن))

جدول ۷. اختلاف اندازه DIF در میان دو گروه از دختران و پسران



جدول ۷. اختلاف مقادیر t-value در میان دو گروه از دختران و پسران



#### ۵- بحث و نتیجه گیری

پژوهش حاضر با هدف بررسی هم‌کنشی توانایی‌های فردی و دشواری آیتم موجود در آزمون‌ها با ریسک و مقیاس جمعیتی بالا در امتحان ورودی دانشگاه در ایران (NUEEFL) انجام شده است. به‌طور خاص، نتایج حاصل از مدل راش و آنالیز DIF مقایسه شد تا بررسی شود آیا شواهدی از تفاوت در عملکرد آیتم‌ها در تجزیه و تحلیل داده‌ها یافت می‌شود.

آمار توصیفی نشان داد که در آزمون کلی اختلاف معنی‌داری وجود دارد. از این رو، فرضیهٔ برآزش داده‌ها با مدل راش پشتیبانی نشد. نتایج برآزش داده‌ها با مدل نشان داد که ۲۶ سوال (شامل سوال‌های ۱۵۵، ۱۲۶، ۱۳۷، ۱۰۵، ۱۱۸، ۱۶۶، ۱۰۱، ۱۰۳، ۱۵۸، ۱۱۱، ۱۱۵، ۱۳۳، ۱۲۱، ۱۶۷، ۱۰۹، ۱۲۸، ۱۲۲، ۱۵۶، ۹۹، ۸۴، ۱۵۳، ۱۴۹، ۱۰۸، ۱۶۰، ۹۱ و ۷۹) از مجموع ۹۵ مورد در محدوده قابل قبول ۰٫۷۰ تا ۱٫۳۰ قرار نداشتند. بنابراین بسیاری از سوال‌های این آزمون، با مدل برآزش همخوانی نداشتند.

با توجه به نتایج بدست آمده از تحقیق، همانطور در جدول فوق مشاهده کردید از بین ۹۵ آیتم، ۴۰ آیتم دارای اختلاف معنی‌دار میان دو گروه دختران و پسران می‌باشند، بنابراین می‌توان ادعا نمود که تغییرناپذیری سئوال‌ات نسبت به جنسیت پاسخ دهندگان مورد قبول واقع نمی‌گردد و فرضیه صفر که بیان می‌کند؛ جنسیت شرکت‌کنندگان یک منبع DIF در آزمون NUEEFL محسوب نمی‌شود، رد می‌شود.

همچنین با توجه به اهمیت DIF در مدل راش در بین گروه جنسیتی، NUEEFL به نظر نمی‌رسد یک آزمون فاقد سوگیری‌های جنسی و فردی نباشد. بنابراین نتیجه‌گیری می‌شود که در پاسخ به آزمون NUEEFL، اختلاف معنی‌داری بین دو گروه زنان و مردان وجود دارد و آزمون NUEEFL برای همه شرکت‌کنندگان مرد و زن عادلانه نیست.



بررسی‌های آماری در مورد یکنواختی آزمون NUEEFL با بهره‌مندی از کالیبراسیون و پیمایش آزمون مدل راش در نرم افزار وینستپ نشان داد که درصد قابل توجهی از آیت‌ها با مدل برازش نداشتند. آزمون بعدیت از طریق تجزیه و تحلیل مولفه اصلی (PCA) بر روی داده‌های خام و باقی مانده‌ها یافت شد. مقدار واریانس توضیح داده شده توسط اجزای مختلف در داده‌ها برابر ۳۴٫۸٪ (مقدار ویژه ۵۰٫۷۰) است که پایین‌تر از میزان تعیین شده است و تک‌بعدی بودن آزمون در کل آزمون ورودی رد شد.

در باره استقلال موضعی، تعدادی آزمون تی-استیودنت انجام شد. بررسی آماره Outfit MnSq نشان داد که مقدار این آماره برای ۲۰ سؤال بیش از ۱٫۳۰ است. با استفاده از تحلیل راش پارامتر دشواری مجدداً براساس این آیت‌ها محاسبه گردید. از تعداد ۵۰۰۰ آزمون تی استیودنت ۲۶۸۰ آزمون معادل ۵۳٫۶ درصد معنی دار شده که از میزان ۵ درصد مورد قبول بسیار بالاتر است. بنابراین شرایط لازم برای تک بعدی بودن و استقلال موضعی سئوال‌ات برقرار نیست و فرضیه استقلال موضعی و بعدیت در کل آزمون NUEEFL پذیرفته نمی‌باشد.

در این راستا، در NUEEFL عملکرد تک به تک افراد باید مورد بررسی قرار گیرد. پیشنهاد می‌شود که شرکت کنندگان می‌بایست بر اساس نقاط ضعف و آسیب‌پذیرشان که از طریق آزمون ورودی دانشگاه مشخص شده است، دست‌بندی شوند. پس از پذیرش در دانشگاه‌ها، آسیب‌پذیری دانشجویی در هر یک از زیربخش‌های آزمون باید به دانشگاه مورد نظر گزارش شود. مشخصاً، این گزارش ارزشمند می‌تواند به دانش آموزان و اساتید کمک کند تا دانشگاه‌ها پس از دریافت گزارش ضعف در هر بخش، به ایجاد دوره‌های پیش‌نیاز انگلیسی قبل از ورود به واحدهای تخصصی دانشجویان بپردازند.

لازم به یادآوری است که در پژوهش حاضر تنها مهارت خواندن که عمدتاً بر روی درک مطلب، واژگان و دستور زبان تمرکز دارد، مورد بررسی قرار گرفت. شایان ذکر است که این مطالعه به بررسی مهارت شنیداری، نوشتن و صحبت کردن نپرداخته است. به عنوان مثال، در مورد درک شنیداری، زنان نسبت به مردان برتری داشتند

در باره استقلال موضعی، تعدادی آزمون تی-استیودنت انجام شد. بررسی آماره Outfit MnSq نشان داد که مقدار این آماره برای ۲۰ سؤال بیش از ۱٫۳۰ است. با استفاده از تحلیل راش پارامتر دشواری مجدداً براساس این آیت‌ها محاسبه گردید. از تعداد ۵۰۰۰ آزمون تی استیودنت ۲۶۸۰ آزمون معادل ۵۳٫۶ درصد معنی دار شده که از میزان ۵ درصد مورد قبول بسیار بالاتر است. بنابراین شرایط لازم برای تک بعدی بودن و استقلال موضعی سئوال‌ات برقرار نیست و فرضیه استقلال موضعی و بعدیت در کل آزمون NUEEFL پذیرفته نمی‌باشد.

در باره استقلال موضعی، تعدادی آزمون تی-استیودنت انجام شد. بررسی آماره Outfit MnSq نشان داد که مقدار این آماره برای ۲۰ سؤال بیش از ۱٫۳۰ است. با استفاده از تحلیل راش پارامتر دشواری مجدداً براساس این آیت‌ها محاسبه گردید. از تعداد ۵۰۰۰ آزمون تی استیودنت ۲۶۸۰ آزمون معادل ۵۳٫۶ درصد معنی دار شده که از میزان ۵ درصد مورد قبول بسیار بالاتر است. بنابراین شرایط لازم برای تک بعدی بودن و استقلال موضعی سئوال‌ات برقرار نیست و فرضیه استقلال موضعی و بعدیت در کل آزمون NUEEFL پذیرفته نمی‌باشد.

آنالیز DIF احتمال دیگری را برای تصدیق آیت‌های تست در گروه‌های جنسیتی تایید کرد. بر اساس نتایج به دست آمده در آنالیز DIF، قابل تفسیر است که از میان ۹۵ آیت، ۴۰ مورد آیت‌های با نمایه DIF مشخص و نشان داده شده است. این نشان می‌دهد که نمره‌های آزمون NUEEFL از واریانس بی ربط به ساختار خالی نیست. از این رو از استدلال و بحث راجع به اعتبار ساختاری (Construct Validity) را پشتیبانی نمی‌کند.

افزون بر این، یک علاقه پیوسته و همیشگی در بین محققان این گستره در مقایسه گروه‌هایی با فرهنگ‌ها، قوم‌ها یا جنسیت‌های مختلف وجود دارد. همچنین، تعصبات جنسیتی و/یا قومی می‌تواند روی یک یا چند گروه تاثیر منفی داشته باشد و به گونه‌ای که ساختار بی‌ارتباط با آزمون ایجاد کنند. در حقیقت، طراحان سوال و برگزارکننده‌های آزمون‌های سرنوشت ساز در بعد وسیع سعی بر ساخت و ایجاد یک آزمون کاملاً عادلانه دارند. در حالی که به دلیل کمبود مطالعات و بررسی‌های آماری و علمی پیرامون آزمون

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., Vol. 4, pp. 221-256). Westport, CT: American Council on Education & Praeger.

Camilli, G., & Penfield, D. A. (1997). Variance estimation for differential test functioning based on the Mantel-Haenszel log-odds ratio. *Journal of Educational Measurement*, 34, 123-139.

Cole, N. S. (1997). *The ETS gender study: How females and males perform in educational settings*. Princeton, NJ: Educational Testing Service.

DeMars, C. (2010). *Item Response Theory: Understanding statistics measurement*. Oxford, UK: Oxford University Press.

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Furr, M. R., & Bacharach, V. R. (2007). *Psychometrics: An Introduction*. Thousand Oaks, CA: SAGE.

Geranpayeh, A., & Kunnan, A. J. (2007). Differential Item Functioning in terms of age in the Certificate in Advanced English Examination. *Language Assessment Quarterly*, 4, 190-222.

Holland, P. W., & Wainer, H. E. (2012). *Differential item functioning*. London, UK: Routledge.

Karami, H. (2015). A closer look at the validity of the University Entrance Exam: Dimensionality and generalizability. (Unpublished Ph.D dissertation, University of Tehran).

Linacre, J. M. (1991-2006). A user's guide to Winsteps® Ministep Rasch-model computer programs. Retrieved January, 10, 2007, from <http://www.winsteps.com/aftp/winsteps.pdf>.

Linacre, J. M. (2012). A user's guide to <https://www.winsteps.com/winman/copyright.htm>

Linacre, J. M. (2016). Winsteps® (Version 3.92.1) [Computer Software]. Beaverton, OR:

(بویل، ۱۹۸۷؛ کول، ۱۹۹۷) (Boyle, 1987; Cole, 1997). این نکته زمینه مطالعه در آن موارد با استفاده از شیوه آماری مطرح شده در این تحقیق را برای افراد علاقه‌مند پیشنهاد می‌کند.

همچنین، تجزیه و تحلیل DIF در این مطالعه نشان داد که بین مردان و زنان عملکرد افتراقی آیت‌ها وجود دارد که یافته‌های این مطالعه متفاوت از نتایج رایان و بکمن (۱۹۹۲) (Ryan & Bachman, 1992) است که هیچ تفاوت جنسیتی در میان هیچ یک از زیر آزمون‌های TOEFL نیافته بودند.

در پایان باید اذعان داشت که سهم عمده این تحقیق مربوط به گستره آزمون‌سازی زبان است. شایسته است که برای چنین آزمون مهم و سرنوشت‌سازی در سطح ملی با ارائه شواهد تجربی برای تفسیر آیت‌ها و نمره‌های NUEEFL از طریق یک تحقیق جامع در رابطه با برازش آیت‌ها، بررسی بعدیت و یکنواختی و یافتن مواردی که باعث ایجاد آیت‌های متعصبانه و سوگیر می‌باشد انجام شود. ارزیابی ساختارها و بررسی عادلانه بودن آزمون، به‌ویژه برای امتحان ورود به دانشگاه در ایران امری ضروری و مهم است. زیرا در حال حاضر این آزمون ملی، تنها راه برای ارزیابی دانش‌آموزان پس از پایان مقطع دبیرستان و تعیین سطح دانش آن‌ها برای ورود به مقطع جدید و بالاتر است که توسط مرجع رسمی کشور، سازمان سنجش (National Organization for Educational Testing (NOET) (NOET) مورد استفاده قرار می‌گیرد.

## Refererance

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 57, No. 1., pp. 289-300

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Dordrecht: Springer Science, and Business Media.

Boyle, J. (1987). Sex differences in listening vocabulary. *Language Learning*, 37(2), 273-284.

Sireci, S. G., & Rios, J. A., (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation*, 19, (2–30), 170–187, doi.org/10.1080/13803611.2013.767621.

Wale, C. M. (2013). Evaluation of the effect of a digital mathematics game on academic achievement (Doctoral dissertation, University of Northern Colorado).

Wiberg, M. (2007). Measuring and detecting differential item functioning in criterion-referenced licensing test: A theoretic comparison of methods. *Educational Measurement*, technical report N. 2.

Zand Scholten, A. (2011). Admissible statistics from a latent variable perspective. *The Institutional Repository of the University of Amsterdam (UvA)*, 29-46.

Winsteps.com. Retrieved from <http://www.winsteps.com/>

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models. Quantitative applications in the social sciences*. Thousand Oaks, CA: SAGE.

Pae, H. (2011). *Differential item functioning and unidimensionality in the Pearson Test of English Academic*. Pearson Education Ltd.

Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 367-388). Hillsdale, NJ: Erlbaum.

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

Scheuneman, J. D., & Bleistein, C. A. (1989). A consumer's guide to statistics for identifying differential item functioning. *Applied Measurement in Education*, 2, 255-275.