



## 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 سوگیری در سوالات درک مطلب آزمون مقطع دکترای رشته زبان انگلیسی تحت سنجش تشخیصی شناختی

#### مسعود گرامی پور

استادیار سنجش و اندازه گیری،  
دانشگاه خوارزمی، تهران، ایران



مسعود گرامی پور دکتری ارزیابی و سنجش می باشد. او به عنوان استادیار در گروه مطالعات درسی دانشگاه خوارزمی از سال ۱۳۹۰ تاکنون فعالیت می کند.

Email: [mgramipour@khu.ac.ir](mailto:mgramipour@khu.ac.ir)

#### حمید مرعشی

دانشیار آموزش زبان انگلیسی،  
دانشکده زبانهای خارجی، دانشگاه  
آزاد اسلامی تهران مرکزی،  
تهران، ایران  
(نویسنده مسئول)



حمید مرعشی دانشیار آموزش زبان انگلیسی، دانشگاه آزاد اسلامی واحد تهران مرکزی و سردبیر مجله زبان و ترجمه است. وی در مجلات TESOL و مجله یادگیری زبان منتشر کرده است.

Email: [hamid.marashi@iauctb.ac.ir](mailto:hamid.marashi@iauctb.ac.ir)

#### مسعود سیری

استادیار آموزش زبان انگلیسی،  
دانشکده زبانهای خارجی،  
دانشگاه آزاد اسلامی علوم و  
تحقیقات، تهران، ایران



مسعود سیری دکتری آموزش زبان انگلیسی می باشد. هم اکنون وی عضو هیئت علمی تمام وقت در واحد علوم و تحقیقات دانشگاه آزاد اسلامی تهران است.

Email: [m.siyari@srbiau.ac.ir](mailto:m.siyari@srbiau.ac.ir)

#### نیلوفر شهیمیرزادی

دانشجوی دکتری آموزش زبان  
انگلیسی، دانشکده زبانهای  
خارجی، دانشگاه آزاد اسلامی  
تهران مرکزی، تهران، ایران



نیلوفر شهیمیرزادی دانشجوی دکتری آموزش زبان انگلیسی دانشگاه آزاد اسلامی واحد مرکزی تهران می باشد. رساله دکتری وی در خبرنامه ILTA 2019 به عنوان رساله دانشجویی برتر انتخاب شده است.

Email: [Niloufar\\_shahmirzadi83@yahoo.com](mailto:Niloufar_shahmirzadi83@yahoo.com)

#### چکیده

#### اطلاعات مقاله

در چند دهه اخیر، بررسی سطح توانایی در آزمونهای سرنوشت ساز از دیدگاه سنجش تشخیصی شناختی مورد توجه محققان ارزشیابی و سنجش و اندازه گیری قرار گرفته است. مدل های تشخیصی شناختی میزان تسلط و عدم تسلط آزمودنی ها را با ویژگی های چندگانه مورد بررسی قرار می دهد. بدین منظور، هدف پژوهش حاضر بررسی کنش افتراقی سوال و کنش افتراقی خصیصه در آزمون ورودی زبان عمومی مقطع دکترای زبان انگلیسی است. بنابراین، ۳۲۲۰ آزمودنی دختر و پسر از جامعه آماری شرکت کننده در آزمون سازمان سنجش کشور به صورت تصادفی انتخاب شدند. در مدل اکتشافی تحلیل داده ها، مراحل کیفی و کمی به ترتیب اجرا شد. سپس از ماتریس کیو ساخته شده به وسیله پروتکل تفکر با صدای بلند استفاده شد تا داده ها با استفاده از بسته آر استودیو تجزیه و تحلیل شوند. مدل جی دینا تحت سنجش تشخیصی شناختی قرار گرفته و نتایج بدست آمده از کنش افتراقی سوال، سوگیری در سوالات را گزارش داده است.

تاریخ ارسال: ۹۸/۰۳/۰۷

تاریخ پذیرش: ۹۸/۰۵/۰۶

تاریخ انتشار: ۹۸/۱۲/۲۵

نوع مقاله: علمی پژوهشی

#### کلید واژگان:

خواندن و درک مطلب، کنش

افتراقی سوال، کنش افتراقی

مهارت، سنجش تشخیصی

شناختی، آزمون ورودی دکتری

کلیه حقوق محفوظ است ۱۳۹۹

شناسه دیجیتال [10.22059/jflr.2019.282328.634:doi](https://doi.org/10.22059/jflr.2019.282328.634)

شهیمیرزادی، نیلوفر، سیری، مسعود، مرعشی، حمید، گرامی پور، مسعود. (۱۳۹۹). بررسی سوگیری در سوالات درک مطلب آزمون مقطع دکترای رشته زبان انگلیسی تحت سنجش تشخیصی شناختی. پژوهشهای زبانشناختی در زبانهای خارجی، ۱۰(۱)، ۱۵۲-۱۵۵. doi: 10.22059/jflr.2019.282328.634



## 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)



### Test Fairness Analysis in Reading Comprehension PhD Nationwide Admission Test Items under CDA

**Niloufar Shahmirzadi**  
Department of Foreign  
Languages, Central Tehran  
Branch, Islamic Azad  
University, Tehran, Iran

**Masood Siyyari**  
Department of Foreign  
Languages, Tehran  
Science and Research  
Branch, Islamic Azad  
University, Tehran, Iran

**Hamid Marashi**  
Department of Foreign  
Languages, Central  
Tehran Branch, Islamic  
Azad University, Tehran,  
Iran  
(corresponding author)

**Masoud Geramipour**  
Kharazmi University, Tehran,  
Iran



Niloufar Shahmirzadi is a Ph.D. candidate of Applied Linguistics from Islamic Azad University, Central Tehran. Her area of research lies in Language Assessment. She has published in ILTA Newsletter 2019.

Email: [niloufar\\_shahmirzadi83@yahoo.com](mailto:niloufar_shahmirzadi83@yahoo.com)

Masood Siyyari is Ph.D. in Applied Linguistics. Currently, he is a full-time faculty member at Science and Research Branch of Islamic Azad University, Tehran. His interest is in Language Assessment.

Email: [m.siyyari@srbiau.ac.ir](mailto:m.siyyari@srbiau.ac.ir)

Hamid Marashi is Associate Professor of Applied Linguistics, Islamic Azad University, Central Tehran and Editor-in-Chief of the Journal of Language and Translation. His publications are TESOL and Language Learning Journals.

Email: [hamid.marashi@iauctb.ac.ir](mailto:hamid.marashi@iauctb.ac.ir)

Masoud Geramipour is Ph.D. in Assessment and Measurement. Currently, he is an Assistant Professor in the Department of Curriculum Studies, at Kharazmi University. He has published some articles.

Email: [mgramipour@khu.ac.ir](mailto:mgramipour@khu.ac.ir)

#### ARTICLE INFO

##### Article history:

Received 28th, May, 2019

Accepted 28th, July, 2019

Available online April 2020

Article Type Research article

##### Keywords:

Reading Comprehension, DIF, DAF, CDA, PhD National Admission Exam

#### ABSTRACT

During the past few decades, documentation of test takers' proficiency level has been accomplished through large-scale assessments most importantly Cognitive Diagnostic Assessment (CDA) in order to provide skills mastery profile of test takers in fine-grained detailed information, and to ascertain multi-diagnostic mastery or non-mastery classifications. Accordingly, the present study attempts to scrutinize reading comprehension test items of a high-stakes test under CDA. To delve into this issue, Differential Attribute Functioning (DAF) was used to detect the probability of mastery of attributes among test takers, and Differential Item Functioning (DIF) was applied to show item performance among different candidates in terms of gender. Thus, the participants of this study were 3220 females and males attending PhD national admission test in Iran. Through adopting sequential exploratory mixed method design, GDINA model was run by the application of R-studio package. Results of the study revealed that test items suspected DIF against female. In the end, the findings of this study were discussed in light of their implications for language testing community to perceive potential social harm which derived from biased test items in PhD national admission exams.

DOI: [10.22059/jflr.2019.282328.634](https://doi.org/10.22059/jflr.2019.282328.634)

© 2020 All rights reserved.

Shahmirzadi, Niloufar, Siyyari, Masood, Marashi, Hamid, Geramipour, Masoud. (2020). Test Fairness Analysis in Reading Comprehension PhD Nationwide Admission Test Items under CDA Journal of Foreign Language Research, 10 (1), 222-155. doi: [10.22059/jflr.2019.282328.634](https://doi.org/10.22059/jflr.2019.282328.634)

## ۱. مقدمه‌ای بر سنجش تشخیصی شناختی

در سال‌های اخیر آزمون‌های سرنوشت ساز (High-Stakes Tests) با رویکرد بهبود یادگیری و رسیدن به موفقیت مورد توجه بسیاری از پژوهشگران قرار گرفته است، زیرا تصمیمات حاصله از آزمون‌های استاندارد در مقیاس کلان پیامدهای مهم "شخصی، اجتماعی و سیاسی" به همراه خواهد داشت (پنفلد و کامیلی، ۲۰۰۷، ۱۲۵). خصوصاً افزایش کیفیت آزمون‌ها در مقاطع تحصیلات تکمیلی، باعث رشد فرهنگی، اجتماعی و اقتصادی می‌شود (نیستانی، ۱۳۹۱). در نتیجه، محققان علم روان‌سنجی و سنجش آموزش در صدد برآمدن تا با رویکردهای نوین موضوع ارزشیابی آموزشی را بازنگری کنند. همبلتون و همکاران (۱۹۹۱) معتقدند، آزمون‌های حرفه‌ای در مقیاس کلان تعیین‌کننده سرنوشت شرکت‌کنندگان در آزمون می‌باشند که لازم است طوری طراحی شوند تا عاری از هرگونه سوگیری باشند.

در ارزشیابی تشخیصی شناختی نه تنها نمره آزمودنی بررسی می‌شود، بلکه با رویکردی جدید مهارت‌های (Skills or Attributes) خاص فراگرفته یا فرانگرفته مورد سنجش قرار می‌گیرند (لیتون و گی ریل، ۲۰۰۷) تا اطلاعات مفیدی را برای حل مشکلات و نشان دادن نقاط قوت و ضعف آزمودنی‌ها تفسیر کنند. این موضوع بر خلاف مدل‌های نظریه سوال پاسخ می‌باشد، چرا که در مدل‌های سنتی اطلاعات دقیقی دربارهٔ نقاط قوت و ضعف آزمودنی‌ها گزارش نمی‌شود و تنها مقایسه و رتبه‌بندی فراگیران گزارش می‌شود (سنو و لومان، ۱۹۸۹).

به بیان دیگر، رویکرد سنجش تشخیصی شناختی (Cognitive Diagnostic Assessment - CDA) اطلاعات دقیق‌تر و عادلانه‌ای را درباره خصیصه‌های مکنون فراگرفته یا نگرفته شده آزمودنی‌ها و انتقال مفاهیم پیچیده ذهنی به تفکیک نشان می‌دهد (هو، دلاتوره و نانداکومار، ۲۰۱۴). در واقع سنجش تشخیصی شناختی نمایی از میزان تسلط یا عدم تسلط (Skill Mastery or Non-Mastery Profile) آزمودنی‌های مورد مطالعه، در مورد وضعیت تسلط آن‌ها در هر مهارت را فراهم می‌کند. هم چنین صفات زیربنایی مورد نیاز برای هر سوال را نشان داده تا بتوان با تدوین و برگزاری برنامه‌های آموزشی جبرانی موضوعات مورد اشکال و سوگیری شده را برطرف کرد. حال با در نظر گرفتن اهمیت سنجش تشخیصی شناختی، ضروری است تا پژوهش‌های بیشتری در مورد کنش افتراقی سوال از دیدگاه

تشخیصی شناختی انجام شود.

## ۲. کنش افتراقی سوال

در بستر سنجش تشخیصی شناختی، مدل‌های مختلفی مطرح می‌شود که در این میان کنش افتراقی سوال (Differential Item Functioning - DIF) به دلیل جدید بودن این روش سنجش کمتر مورد بررسی قرار گرفته است. کنش افتراقی سوال زمانی مطرح می‌شود که آزمودنی‌ها از گروه‌های مختلف و با تسلط یکسان احتمال متفاوت در پاسخ صحیح به سوالات را دارند. به منظور برطرف کردن این مشکل، با تحلیل کنش افتراقی سوال بررسی تغییرپذیری یا عدم تغییرپذیری سازه فراهم می‌شود (زومبو، ۲۰۰۷). به اعتقاد هو، دولا توره و نانداکومار (۲۰۱۴)، لی و وانگ (۲۰۱۵) در کنش افتراقی سوال تحت مدل‌های تشخیصی شناختی، رابطه بین هر سوال با صفات مرتبط در گروه‌های آزمودنی با پیشینه متفاوت محاسبه می‌شود، تا سوالات دارای سوگیری که منصفانه بودن آزمون را زیر سوال برده‌اند، مشخص شود.

با توجه به اهمیت کارکرد افتراقی سوال، پژوهش‌هایی در زمینه اندازه‌گیری با روش‌های آماری مختلف از دیدگاه نظریه کلاسیک آزمون (Classical Test Theory - CTT) و نظریه سوال پاسخ (Item Response Theory - IRT) صورت گرفته است. برای مثال در نظریه کلاسیک آزمون، روش منتل (Mantel) (منتل، ۱۹۶۳)، روش منتل-هنزل (Mantel-Haenszel) (منتل-هنزل، ۱۹۵۹، هولاند و تایر، ۱۹۸۸)، روش منتل - هنتزل تعمیم یافته (Generalized Dichotomous Mantel-Haenszel Method for Items) برای داده‌های چند ارزشی (منتل و هنتزل، ۱۹۵۹)، آزمون سوگیری همزمان سوال (Simultaneous Item Bias Test - SIBTEST) (شیلی و استات، ۱۹۹۳a) شیلی و استات (b، ۱۹۹۳) و روش رگرسیون لجستیک (Logistic Regression) (سوامیناتان و راجرز، ۱۹۹۰) مطرح می‌شود. در روش نظریه سوال پاسخ، روش آزمون نسبت درست نمایی (Likelihood Ratio Test)، آزمون والد (Wald Test) (لرد، ۱۹۸۰؛ تیسن، استاینبرگ و واینر، ۱۹۸۸) و روش شاخص‌های چندگانه و علل چندگانه (Multiple Indicators Multiple Causes Method) (فینچ، ۲۰۰۵) با مفروضات متفاوت مورد استفاده قرار گرفته شده است.

حال با توجه به جدید بودن مدل تشخیصی شناختی، ضروری است تا کنش افتراقی سوال تحت مدل تشخیصی

شناختی با آزمون والد محاسبه شود. شایان ذکر است که در این پژوهش تحت مدل جی دینا کنش افتراقی سوال بررسی شده است.

### ۳. پیشینه پژوهش

به اعتقاد دی بلو، روسوس و استات (۲۰۰۷)، راپ و تمپلین (۲۰۰۸)، راپ، تمپلین و هسن (۲۰۱۰)، در سال‌های اخیر مطالعاتی در مورد طبقه‌بندی مبتنی بر مدل‌های تشخیصی شناختی با در نظر گرفتن پایایی (Reliability) کمی صورت گرفته است تا به سوال دیرینه در حوزه سنجش و اندازه‌گیری پاسخ داده شود. این سوال کنش افتراقی سوال را در سنجش تشخیصی شناختی مطرح می‌کند.

کنش افتراقی سوال، زمانی دیده می‌شود که آزمودنی‌ها با یک سطح توانایی در شرایط یکسان احتمال متفاوتی در پاسخ به سوال داشته باشد (سوامیناتان و راجرز، ۱۹۹۰). به بیان دیگر، عدم برقراری عدالت در طراحی سوالات می‌تواند سوگیری متفاوتی نسبت به هر آزمودنی نشان دهد که در نهایت روایی (Validity) آزمون را تهدید می‌کند. زومبو (۱۹۹۹، ۶) "بر سودار نبودن آزمون‌ها با جامعه آماری بالا" تاکید می‌کند تا عدالت در نشان دادن توانایی آزمودنی‌ها در سطح ملی و با جامعه آماری بالا اجرا شود. به بیان روشن‌تر، سودار نبودن سوالات از جنبه‌های مختلف جنسیتی، اجتماعی، اقتصادی، فرهنگی و نژادی (کول و زیکی، ۲۰۰۱؛ مک نامارا و روور، ۲۰۰۶؛ به نقل از سانگ، چنگ و کلینبرگ، ۲۰۱۵) می‌بایست قبل از برگزاری آزمون مورد توجه دقیق قرار گیرد. حال نکته دیگری که در آزمون‌های استاندارد لازم است در نظر گرفته شود وجود کنش افتراقی سوال در آزمون‌ها، با توجه به پیشینه پژوهشات انجام شده است. به عنوان مثال سانگ، چنج و کلینگرگ (۲۰۱۵) وجود تبعیض جنسیتی و تاثیر میزان پیشینه تحصیلات در آزمون‌های استاندارد تحصیلات تکمیلی را با توجه به عملکرد آزمودنی‌ها مورد بررسی قرار داده‌اند. هم چنین، گزارش پژوهش انجام شده توسط امیران، علوی و فیدلگو (۲۰۱۴) حاکی از آن است که ۲۸٪ سوال‌های آزمون دارای کنش افتراقی جنسیتی می‌باشند. براتی و احمدی (۲۰۱۰) با رویکرد نظریه پاسخ به سوال به بررسی کنش افتراقی در سوالات آزمون ورودی دانشگاه سراسری پرداختند. قابل توجه آن است که در آزمون‌هایی با سوالات محتوایی علوم سخت، جامعه آماری دختران و با مضامین علوم انسانی جامعه آماری پسران عموماً مورد سوگیری قرار می‌گیرند.

شایان ذکر است که در آزمون بین‌المللی

تیمز (TIMESS) نیز بیش از ۲۰٪ سوالات هر دفترچه کنش افتراقی را نشان می‌دهند (دودین و انابی، ۲۰۰۸). بارنز و ولز (۲۰۰۹) نقش جنسیت و نژاد را در کنش افتراقی بررسی نمودند. یونگ، مورگان، ریپنسکی، استینبرگ و وانگ (۲۰۱۳) کنش افتراقی در سوالات آزمون تافل را مورد پژوهش قرار دادند و در گزارشی مطرح کردند که جامعه آماری مردان از سوالاتی با محتوای ورزشی درک بهتری داشتند. هم چنین، شاموگام و سالان (۲۰۱۴) نقش زبان مادری را در کنش افتراقی سوال آزمون موثر دانستند.

از سوی دیگر، کنش افتراقی خصیصه (Differential Attribute Functioning DAF) در مدل تشخیصی شناختی به منظور بررسی نقش تغییرپذیری خصیصه‌ها در پاسخ به سوال مورد سنجش قرار گرفته شده است. لی (۲۰۰۸، ۲۰۱۱) در پژوهشی بر لزوم یکسان بودن کنش افتراقی خصیصه در سوالات تاکید دارد چرا که کنش افتراقی خصیصه تاثیر گروهی از خصیصه‌ها را در مجموعه ای از سوالات نشان می‌دهد. علاوه بر آن، رنجبران و علوی (۲۰۱۶) نیز بر ضرورت یادگیری خصیصه‌ها در آزمون مهارت خواندن و درک مطلب زبان انگلیسی تاکید کرده‌اند. به منظور نیل به این هدف، با طراحی ماتریس کیو می‌توان خصیصه‌های مکنون در هر سوال را مورد بررسی قرار داد.

### ۴. طراحی ماتریس کیو در شناسایی خصیصه‌ها

در مدل‌های تشخیصی شناختی "رابطه بین خصیصه‌های شناختی و سوالات" با طراحی ماتریس کیو (Q-matrix) (ینگ، ۲۰۰۹، ۲۱۴) نشان داده می‌شوند. این ارتباط شامل ماتریس دوگانه  $[i \times k]$  است. در این ماتریس، خصیصه‌های صفر نشان دهنده عدم وجود و خصیصه‌های یک بیان‌نگر وجود آن خصیصه در سوال مورد نظر است. تاتسوکا (۱۹۸۳) معتقد است در مدل‌های تشخیصی شناختی مشخص کردن ارتباط سوالات با خصیصه مکنون حائز اهمیت است. دولاتوره (۲۰۰۹، ۲) تاکید "بر طراحی ماتریس کیو با هدف مشخص است. تاتسوکا (۱۹۸۳) معتقد است در مدل‌های تشخیصی شناختی مشخص کردن ارتباط سوالات با خصیصه مکنون حائز اهمیت است. دولاتوره (۲۰۰۹، ۲) تاکید "بر طراحی ماتریس کیو با هدف مشخص کردن دیدگاه شناختی هر سوال دارد." فالماگن و دوگان (۱۹۸۸) ماتریس طراحی شده ابتدایی توسط تاتسوکا (۱۹۹۰) را بسط داده و در طرحی نشان دادند:

$$\left\{ \begin{array}{l} ۱ \text{ وجود خصیصه } k \text{ در سوال} \end{array} \right\}$$

## qik

• عدم وجود خصیصه k در سوال i

تا در تدوین ماتریس کیو شرکت کنند.

### ۲-۱-۵ آزمودنی‌های کمی

نمونه پژوهش از بانک داده‌های خام سازمان سنجش و آموزش کشور تهیه شد. به منظور تحلیل داده در مرحله کمی ۳۲۲۰ نمونه از جامعه آماری به صورت تصادفی انتخاب شدند. این تعداد ۱۹۴۵ آزمودنی دختر و ۱۲۷۵ آزمودنی پسر را شامل می‌شود که بین متوسط سنی ۲۵ تا ۴۵ می‌باشند. شایان ذکر است که آزمودنی‌ها، داوطلب شرکت در آزمون مقطع دکترای زبان انگلیسی بودند که می‌بایست سوالات دفترچه عمومی زبان انگلیسی سال ۱۳۹۵ را در ۶۰ دقیقه پاسخ دهند.

### ۲-۵ ابزار اندازه‌گیری

#### ۱-۲-۵ ابزار اندازه‌گیری کیفی

به منظور تشخیص دقیق ویژگی‌ها در پاسخ به سوالات بخش خواندن و درک مطلب زبان انگلیسی در سنجش شناختی تشخیصی، از مدل گنو و راجرز (۲۰۱۰) و ینگ (۲۰۰۹) استفاده شد. در این خصوص، از دانشجویان دعوت شد تا در پروتکل کلامی با صدای بلند شرکت کنند. آن‌ها می‌بایست پس از خواندن متن درک مطلب و سوالات، به پرسش‌نامه تهیه شده پاسخ دهند. در پرسش‌نامه مذکور، هر سوال به‌طور جداگانه با در نظر گرفتن پنج خصیصه مورد تجزیه و تحلیل قرار گرفتند و سپس از هر دانشجوی شرکت کننده خواسته شد تا نظرات و افکار خود را در مورد هر سوال در حین پاسخ به سوالات بلند بیان کنند. در مرحله آخر، اساتید مربوطه نیز نظر نهایی خود را در مورد هر سوال و خصیصه‌های ضروری در پاسخ به آن سوال در پروتکل کلامی با صدای بلند مطرح کردند. جدول ۱ ویژگی‌های بکار برده شده در این پژوهش را نشان می‌دهد.

در این ماتریس اعداد صفر و یک با ضریب  $k, i$  نشان داده می‌شوند. اما آنچه در طراحی ماتریس کیو مورد توجه قرار می‌گیرد، دقت بالا در شناسایی ویژگی‌های مکنون می‌باشد (ینگ، ۲۰۰۵، ۲۰۰۹) چرا که عدم رعایت عینیت در مراحل طراحی ماتریس می‌تواند سوگیری‌های شخصی را در طراحی ماتریس نشان دهد. نکته دیگری که در تدوین ماتریس کیو مورد اهمیت است، روش‌های پیشنهادی (تاتسکو، ۱۹۸۳) از جمله روش مشورت با افراد متخصص، ارزیابی ذهنی و بررسی سوالات آزمون می‌باشد. ینگ (۲۰۰۵) نیز در رساله دکتری خود از روش پروتکل تفکر با صدای بلند (Think Aloud Verbal Protocol Analysis) استفاده کرد. در این روش مبنا ابتدایی، مطالعه پیشینه پژوهش، مطالعه پروتکل تفکر با صدای بلند توسط دانشجویان و متخصصان می‌باشد. سپس ماتریس تدوین شده با استفاده از روش‌های آماری اعتباریابی می‌شود (لی و سوئن، ۲۰۱۳، ۹۷).

با توجه به ضرورت طراحی ماتریس کیو، در پژوهش حاضر ابتدا ماتریس کیو طراحی شده، سپس کنش افتراقی سوال و کنش افتراقی خصیصه در راستای سنجش تشخیصی شناختی مورد پژوهش قرار می‌گیرد. در نیل به این هدف، سوالاتی در بررسی میزان سوگیری سوال و خصیصه مطرح شده است. ۱. آیا کنش افتراقی سوال، سوگیری جنسیتی در سوالات درک مطلب آزمون زبان عمومی ورودی مقطع دکتری دانشگاه سراسری داشته است؟ ۲. آیا کنش افتراقی خصیصه، سوگیری جنسیتی در سوالات درک مطلب آزمون زبان عمومی ورودی مقطع دکتری دانشگاه سراسری داشته است؟

### ۵. روش پژوهش

#### ۱-۵ آزمودنی‌ها

##### ۱-۱-۵ آزمودنی‌های کیفی

در مرحله کیفی، به منظور تدوین ماتریس کیواز ۵ دانشجوی ترم اول مقطع دکترای آموزش زبان انگلیسی و ۸ استاد در رشته آموزش زبان انگلیسی دعوت شد تا در تدوین پروتکل کلامی با تفکر بلند شرکت کنند. دانشجویان بین رده سنی ۲۵ تا ۴۵ بوده و استادان شرکت کننده نیز بین ۲۰ تا ۳۰ سال سابق آموزش در رشته آموزش زبان انگلیسی را داشتند. شایان ذکر است که پژوهش حاضر به روش مهندسی معکوس انجام شد. بدین علت از دانشجویان مقطع دکتری دعوت شد

جدول ۱. خصیصه‌های (مهارت‌های) آزمون خواندن و درک مطلب

| مهارت‌های خواندن   | پژوهشگران                         |
|--|-----------------------------------|
| واژگان (Vocabulary)<br>نحو (Syntax)<br>استخراج اطلاعات متنی<br>(Extracting Explicit Information) | گنو و راجرز (۲۰۱۰)<br>جانگ (۹۹۹۹) |
| ارتباط و ترکیب<br>(Connecting and Synthesizing)  |                                   |

(Making Inference)

و اساتید رشته آموزش زبان انگلیسی در پروتکل تفکر با صدای بلند، داده‌های خام با استفاده از نرم افزار آماری اس پی اس مورد تجزیه و تحلیل قرار گرفت. به منظور بررسی میزان توافق بین دانشجویان و استادان در طراحی ماتریس کیو ابتدایی، در آزمون کاپا میزان توافق نظر محاسبه شد تا نظر هر دو گروه در نظر گرفته شود (جدول ۲).

۲-۲-۵ ابزار اندازه گیری کمی  
ابزار بکار برده شده در بخش کمی، سوالات زبان عمومی آزمون ورودی مقطع دکترای رشته زبان انگلیسی (گرایش‌های مترجمی، ادبیات و آموزش زبان انگلیسی) بوده است. این آزمون در ۳۰ سوال چهار گزینه طراحی شده که سه بخش سوالات لغت، گرامر و درک مطلب را شامل می‌شود. با توجه به هدف پژوهش، بخش خواندن و درک مطلب که دارای ۱۰ سوال می‌باشد، مورد تجزیه و تحلیل آماری قرار گرفت. روش نمره‌گذاری سوال‌های آزمون به صورت دو ارزشی (صفر و یک) می‌باشد. بدین معنا که نمره یک نیاز به آن مهارت‌ها در پاسخ به آن سوال و نمره صفر عدم نیاز به آن مهارت‌ها در پاسخ به آن سوال را نشان می‌دهد.

جدول ۲. آزمون ضریب توافق کاپا

## ۶. تجزیه و تحلیل آماری

در ابتدا به منظور بررسی میزان اتفاق نظر بین دانشجویان

|                            | Value | Asymp.Std.<br>Error <sup>a</sup> | Approx. T <sup>b</sup> | Approx. |
|----------------------------|-------|----------------------------------|------------------------|---------|
| Measure of Agreement Kappa | 1.000 | 0.000                            | 3.162                  | .002    |
| N of Valid Cases           | 10    |                                  |                        |         |

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

اند. هم چنین می‌بایست در محاسبه کنش افتراقی خصیصه، ارتباط یا عدم ارتباط کامل بین ویژگی‌ها بررسی شود (جدول ۳).

با توجه به نتایج به دست آمده از آزمون کاپا، تقریباً توافق کامل بین شرکت‌کنندگان در پروتکل تفکر با صدای بلند وجود دارد ( $k=1.00$ ).

نکته دیگری که می‌بایست در این رابطه مورد توجه قرار گیرد میزان اطمینان از رابطه بین خصیصه‌های بررسی شده در متن خواندن و درک مطلب است. در نتیجه، از آزمون فی جهت بررسی میزان ارتباط بین ویژگی‌ها استفاده شد تا

جدول ۳. آزمون فی در بررسی میزان رابطه بین ویژگی

اطمینان حاصل شود که آزمودنی‌ها از دانش خود و نه از وجود ارتباط بین ویژگی‌ها، در پاسخ به سوالات استفاده کرده

| نتیجه ی | ارتباط وتر | ت ی | نحو | واژگان | خصیصه ه |
|---------|------------|-----|-----|--------|---------|
|---------|------------|-----|-----|--------|---------|

|             |         |         |        |   |   |
|-------------|---------|---------|--------|---|---|
| اژگان       | ۱       |         |        |   |   |
|             | -۰.۰۰۳۴ | ۱       |        |   |   |
| استخراج ت س | ۰.۹۳۹۹  | ۰.۲۴۴۳۲ | ۳      |   |   |
|             | ۰.۵۵۵۴۵ | -۰.۵۱۴۵ | -۰.۰۷۱ | ۱ |   |
|             | ۰       | ۰       | ۰      | ۰ | ۱ |

Note: Bold font indicates strong correlations (> |.70|)

نکته حائز اهمیت دیگر، میزان پایایی بین سوالات آزمون می باشد تا اطمینان حاصل شود که کلیه سوالات به درستی طراحی شده است. در جدول ۴، با استفاده از آزمون کرونیباخ آلفا این میزان مورد اندازه گیری قرار گرفته شد.

با توجه به جدول ۳، رابطه منفی بین واژگان و نحو، نحو و ارتباط و ترکیب، استخراج اطلاعات متنی و ارتباط و ترکیب وجود دارد. هم چنین، رابطه بین واژگان و استخراج اطلاعات متنی، واژگان و ارتباط و ترکیب، نحو و استخراج اطلاعات متنی نشان داده شد. در نهایت ۱/۱۵ (۰.۶۶٪) رابطه قوی، ۱/۱۵ (۰.۶۶٪) رابطه متوسط و ۱۳/۱۵ (۰.۸۶٪) رابطه ضعیف بین خصیصه‌ها دیده شده است.

جدول ۴. آمار پایایی سوالات در سال ۲۰۱۷

| Cronbach's Alpha<br>فای کرو | Cronbach's Alpha Based on Standardized Items<br>آلفای کرونیباخ براساس | No of Item<br>تعداد سوا |
|-----------------------------|---|-------------------------|
| .۷۹۹                        | .۷۱۷  | ۱۰                      |

جدول ۵. آمار مجموع سوالات

| Item | Scale Mean if Item Deleted | Scale Variance if Item Deleted | Corrected Item-Total Correlation | Squared Multiple Correlation | Cronbach's Alpha if Item Deleted |
|------|----------------------------|--------------------------------|----------------------------------|------------------------------|----------------------------------|
| 1    | 1.85                       | 3.753                          | .419                             | .186                         | .702                             |
| 2    | 1.99                       | 4.283                          | .223                             | .061                         | .728                             |
| 3    | 1.93                       | 4.204                          | .199                             | .049                         | .733                             |
| 4    | 1.83                       | 3.632                          | .484                             | .258                         | .691                             |
| 5    | 1.91                       | 3.924                          | .378                             | .161                         | .709                             |
| 6    | 1.91                       | 3.949                          | .354                             | .134                         | .712                             |
| 7    | 1.76                       | 3.520                          | .503                             | .280                         | .687                             |
| 8    | 1.73                       | 3.431                          | .539                             | .324                         | .679                             |
| 9    | 1.85                       | 3.784                          | .401                             | .174                         | .705                             |
| 10   | 1.88                       | 3.954                          | .324                             | .114                         | .717                             |

شده جهت نشان دادن رابطه سوال و مهارت و استفاده در مراحل بعدی آماری طراحی شد (جدول ۶). در ماتریس تدوین شده با در نظر گرفتن این موضوع که اکثر سوالات به حداقل سه مهارت مرتبط شده اند، می توان این ماتریس را از دیدگاه تشخیصی شناختی قابل اعتماد دانست (هارتز، ۲۰۰۲).

جدول ۶. ماتریس کیو خواندن و درک مطلب

با توجه به نتایج بدست آمد، پایایی قابل قبولی بین سوالات آزمون وجود دارد  $0.7 \leq \alpha < 0.8$  ( $\alpha = .75$ ).

پس از تجزیه و تحلیل مقدماتی سوالات آزمون خواندن و درک مطلب و پروتکل تفکر با صدای بلند، ماتریس کیو نهایی

| سوال | واژگان | نحو | استخراج اطلاعات متنی | ارلاطو و ترلاب | نتیجه گیری |
|------|--------|-----|----------------------|----------------|------------|
|------|--------|-----|----------------------|----------------|------------|

|    |   |   |   |   |   |
|----|---|---|---|---|---|
| ۱  | ۱ | ۰ | ۰ | ۱ | ۱ |
| ۲  | ۰ | ۰ | ۰ | ۰ | ۱ |
| ۳  | ۱ | ۰ | ۰ | ۱ | ۱ |
| ۴  | ۰ | ۰ | ۱ | ۱ | ۱ |
| ۵  | ۰ | ۱ | ۱ | ۰ | ۱ |
| ۶  | ۰ | ۰ | ۰ | ۱ | ۱ |
| ۷  | ۱ | ۱ | ۰ | ۰ | ۱ |
| ۸  | ۱ | ۰ | ۰ | ۱ | ۱ |
| ۹  | ۱ | ۱ | ۰ | ۱ | ۱ |
| ۱۰ | ۱ | ۰ | ۰ | ۰ | ۱ |

| آزمون درک مطلب مدل‌جی‌دینا | AIC<br>معیار انتخاب مدل | BIC<br>معیار رفع مشکل مدل انتخابی در رگرسیون | -2Loglike |
|----------------------------|-------------------------|--|-----------|
|                            | ۹۷۹۷۸                   | ۲۸۶۲۲  | -۳۳۸۸۲.۹۵ |

با توجه به جدول ۷، میزان سه شاخص برازش جنسیتی به ترتیب شامل (AIC=27978)، (BIC=28622) و (LRT=-13882.95) می‌شوند که در این میان شاخص AIC کمتر از BIC بوده است و مدل انتخابی جی دینا (GDINA) از برازش خوبی برخوردار است. در مرحله بعدی، با استفاده از نرم افزار آر استودیو (R-Studio) و مدل جی دینا، شاخص‌های مطلق برازش مدل (چن، توره و ژانگ، ۲۰۱۳) محاسبه شد. نتایج بدست آمده که نشان دهنده برازش خوب با مدل می‌باشد عبارتند از  $RMSEA < 0.05$  و  $SRMSR = 0.01$  و  $p < 0.05$

در ماتریس بدست آمده سوال‌های آزمون به صورت دو ارزشی (صفر و یک) نمره گذاری شدند. عدد ۱ مبین خصیصه‌های مکنون ضروری در پاسخ به آن سوال و عدد ۰ بیانگر صفت‌هایی است که در پاسخ به سوال لازم نمی‌باشند. در مرحله بعد، با توجه به ماتریس طراحی شده و با استفاده از مدل جی دینا، شاخص‌های برازش جنسیتی شامل -AIC, BIC, 2Loglike محاسبه شد (جدول ۷).

جدول ۷. شاخص‌های برازش جنسیتی

جدول ۸. شاخص مطلق برازش مدل چند گروهی جی دینا

| Item | Weighted Root Mean Square Deviation (WRMSD) | Bias Corrected WRMSD | Mean Absolute Deviation (MAD) |
|------|---|----------------------|-------------------------------|
| 1    | 0.020298516                                 | 0.018299145          | 0.007125389                   |
| 2    | 0.025607529                                 | 0.024250479          | 0.024268430                   |
| 3    | 0.010265863                                 | 0.008471863          | 0.008477278                   |
| 4    | 0.002319371                                 | 0.000000000          | 0.001288637                   |
| 5    | 0.011604424                                 | 0.008896000          | 0.004955944                   |
| 6    | 0.024123515                                 | 0.022597726          | 0.015487858                   |
| 7    | 0.019159441                                 | 0.016953376          | 0.009473925                   |
| 8    | 0.009413039                                 | 0.005750183          | 0.006060867                   |
| 9    | 0.011623011                                 | 0.008445851          | 0.005657233                   |
| 10   | 0.023077936                                 | 0.021236626          | 0.010880629                   |
| Mean | 0.015749                                    | 0.01349              | 0.009368                      |

جزئیات بیشتر، تحت مدل جی دینا در مهارت خواندن و درک مطلب را نشان می‌دهد. پس از محاسبه شاخص‌های برازش و

هم چنین ارزش  $(Mx^2=4.58)$ ,  $p=1.00$  نیز برازش مدل جی دینا را تایید می‌کند. جدول ۸ شاخص برازش مطلق با



برآورد پارامترها، امکان محاسبه کنش افتراقی سوال فراهم شد (جدول ۹). سپس تحت CDM پکیج آر استودیو، با استفاده از آزمون والد کارکرد افتراقی سوال مورد بررسی قرار گرفت. جدول ۹. کنش افتراقی سوال با آزمون والد

| سوال | Wald Statistic<br>آزمون والد | Adjusted P-Value<br>پی تبیین یافته |
|------|------------------------------|------------------------------------|
| ۱    | ۴۷.۱۱۸۴                      | ۰.۰۰۰**                            |
| ۲    | ۱۲.۲۲۳۷                      | ۰.۱۱۵۵*                            |
| ۳    | ۳۸.۵۸۴۱                      | ۰.۰۰۰۱**                           |
| ۴    | ۷۲.۴۹۱۴                      | ۰.۰۰۰**                            |
| ۵    | ۶۲.۸۱۸۱                      | ۰.۰۰۰**                            |
| ۶    | ۶۲.۶۴۱۶                      | ۰.۰۰۰**                            |
| ۷    | ۸۹.۱۰۴۱                      | ۰.۰۰۰**                            |
| ۸    | ۹۶.۳۲۸۴                      | ۰.۰۰۰**                            |
| ۹    | ۴۸.۵۵۵۵                      | ۰.۰۰۰**                            |
| ۱۰   | ۷۲.۹۹۷۸                      | ۰.۰۰۰**                            |

Note: adjusted p-values based on the Bonferroni correction.

Note: Effect Size shows based on

غیر قابل تایید \*

بزرگ \*\*

ناچیز \*\*\*

نتایج آزمون والد در محاسبه کنش افتراقی سوال میزان پی تبیین یافته را نشان می دهد. اکثریت سوالات از جمله ۱، ۳، ۴، ۵، ۶، ۷، ۸ و ۱۰ دارای سوگیری جنسیتی با میزان ضریب اثر کنش افتراقی بالا گزارش شده است. افزون بر آن، نتایج کنش افتراقی ویژگی برای دختران و پسران با روش متل و هنزل محاسبه شد (جدول ۱۰)

جدول ۱۰. آمار متل هنزل در کنش افتراقی خصیصه

| خصیصه                | MH Chi-Square<br>مجدور کا- متل هنزل | P-Value<br>مقدار پی |
|----------------------|-------------------------------------|---------------------|
| واژگان               | ۰.۴۱۷۷                              | ۰.۶۱۸۶              |
| نحو                  | ۱.۳۳۴۵                              | ۰.۲۸۶۸              |
| استخراج اطلاعات متنی | ۱.۷۷۷۹                              | ۰.۰۱۰۰              |
| ارتباط و ترکیب       | ۰.۰۸۰۴                              | ۰.۷۷۷۷              |
| نتیجه گیری           | ۰.۰۰۴۱                              | ۰.۹۴۸۸              |

با توجه به نتایج بدست آمده از کنش افتراقی سوال، سوگیری خصیصه در میان پنج خصیصه خواندن و درک مطلب مشاهده نشده است.

## ۷. بحث و نتیجه گیری

نتایج آماری حاصل از کنش افتراقی سوال و کنش افتراقی ویژگی در آزمون ورودی مقطع دکترای زبان عمومی انگلیسی تحت مدل جی دینا در سنجش تشخیصی شناختی مورد بررسی قرار گرفت و وجود سوگیری در کنش افتراقی سوال و عدم وجود سوگیری در کنش افتراقی ویژگی ها را نشان می دهد. به اعتقاد رور (۲۰۰۷، ۱۸۴)، در سنجش و اندازه گیری نه تنها سازه های نامربوط مورد بررسی قرار می گیرند، بلکه بررسی "عدالت" در سوالات طراحی شده نیز مورد توجه

می باشد. این موضوع زمانی از اهمیت بیشتری برخوردار می شود که با سرنوشت جامعه آزمودنی ها روبرو هستیم. لیم (۲۰۱۵) نیز تاکید می کند که در ارزشیابی آموزشی و روانشناختی نوین، اطلاعات دقیق تری به منظور یادگیری آزمودنی ها ارائه می شود تا با رعایت عدالت اجتماعی کلیه آزمودنی ها از فرصتی مشابه برای پاسخ گویی به سوالات برخوردار شوند. امبرتسون (۱۹۸۳)، ادامز، ویلسون و ونگ (۱۹۹۷) نیز بر این باورند که به طور متداول در آزمون های سرنوشت ساز به ارائه اطلاعات تشخیصی شناختی در مورد نقاط قوت و ضعف آزمودنی ها توجه نمی شود و توانایی شرکت کنندگان در یک نمره خلاصه می شود. به منظور رفع این مشکل، نکاتی از دیدگاه شناختی می باید، مورد بررسی بیشتر قرار گیرد. ماهیت این اندازه گیری شامل سنجش متغیرهای پنهان در بررسی توانایی و فعالیت آزمودنی ها می شود. در این رابطه، محققان بر این باورند که تعاملات اجتماعی بین افراد و پردازش های شناختی در ذهن آنها می تواند سیستم های پیچیده فکری را به وجود آورد. این نوع تفکر با مجموعه ای وسیع از تعاملات در زندگی حرفه ای و عادی شکل می گیرند و مجموعه این فعالیت ها در محیط های خاص با اهداف و فرهنگ های ویژه و متنوع دیده می شود. در نتیجه، این تنوع باعث می شود تا آزمودنی ها با پیشینه های متفاوت مورد سنجش و ارزشیابی قرار گیرند.

با توجه به موارد مطرح شده از یک سو، و از دیدگاه عملی در نظام سنجش و اندازه گیری از سوی دیگر، خدایی (۱۳۸۸، ۲۰) هدف از تغییرات و اصلاحات در نظام آموزشی برای پذیرش داوطلبان در آزمون های پذیرش دانشجو را "ایجاد شانس برابر برای همه داوطلبان از لحاظ وضعیت فرهنگی، اجتماعی، اقتصادی و اجتماعی" خوانده است. اما محدودیت ظرفیت یا عدم ارائه رشته دانشگاهی برای هر دو گروه جنسیتی می تواند باعث ایجاد سوگیری در سوالات آزمون شود. در نهایت این موضوع سبب ایجاد رقابت شدید بین داوطلبان در مقاطع تحصیلات تکمیلی دانشگاه ها شده و باعث می شود تا عوامل مداخله گر دیگری در پذیرش دانشجو مورد اهمیت قرار گیرند.

در آخر، به اعتقاد پای (۲۰۰۴)، آزمون خواندن و درک مطلب زبان انگلیسی به همراه آزمون ریاضی از اهمیت بالایی در سنجش تشخیصی شناختی آزمودنی ها برخوردارند. پژوهش حاضر، این مهم را از دیدگاه کنش افتراقی سوال و کنش افتراقی خصیصه مورد پژوهش و بررسی قرار داده است. اما آنچه که در این هدف اهمیت به سزایی دارد بهبود بخشیدن در

<http://ensani.ir/fa/article/304325/gender-based-dif-across-the-subject-area-a-study-of-the-iranian-national-university-entrance-exam>

Barnes, B. J., & Wells, C. S. (2009). Differential item functional analysis by gender and race of the national doctoral program survey. *International Journal of Doctoral Studies*, 4, 77–96.

<http://ijds.org/Volume4/IJDSv4p077-096Barnes258.pdf>

Chen, J., Torre, J., & Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, 00(2), 123–140.

<https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-3984.2012.00185>

De La Torre, J. (2009). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, 33, 163–183.

<https://journals.sagepub.com/doi/abs/10.1177/0146621608320523>

DiBello, L. V., Roussos, L., & Stout, W. F. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In: Rao CR, Sinharay S (eds) *Handbook of statistics*, vol 26. Amsterdam, Elsevier, pp 979–1030.

<https://www.sciencedirect.com/science/article/pii/S0169716106260310>

Doudeen, H., & Annabi, H. (2008). Sex-Related Differential Item Functioning (DIF) Analysis of TIMSS. *Educational Sciences*, 55(697).

<https://journals.ju.edu.jo/DirasatHum/article/view/1807>

Embretson (Whitely), S. E. (1983). Construct

طراحی سوالات آزمون‌های ملی می‌باشد. بدین منظور پیشنهاد می‌شود تا پژوهش بیشتری بر روی آزمون‌های کلان تحت دیدگاه تشخیصی شناختی صورت گیرد و سوالات آزمون رشته‌های دیگر دانشگاهی نیز مورد توجه و پژوهش قرار گیرند. هم‌چنین، از متخصصان آزمون سازی دعوت شود تا با طراحی سوالات حرفه‌ای و نه صرفاً دشوار به سنجش و ارزشیابی دقیق‌تری از آموذنی‌ها در مقیاس بزرگ بپردازند. اگر چه به دلیل نبودن سنجش تشخیصی شناختی نیاز است محدودیت تدوین ماتریس کیو استاندارد در نظر گرفته شود چرا که تاکنون در این زمینه روشی استاندارد پیشنهاد نشده است.

#### منابع

خدایی، ابراهیم. (۱۳۸۸). الف. بررسی عوامل مؤثر بر قبولی در آزمون کارشناسی ارشد. فصلنامه پژوهش و برنامه‌ریزی در آموزش عالی. شماره ۵۴، صص ۳۴–۱۹.

<http://journal.irphe.ac.ir>

نیستانی، محمدرضا. (۱۳۹۱). برنامه‌ریزی آموزشی راهبردهای بهبود کیفیت در سطح یک واحد (مدرسه، واحد دانشگاهی و آموزش مجازی) اصفهان: آموخته.

<https://www.adinehbook.com/gp/product/6006465>

012

Adams, R. J., Wilson, M. R., & Wang, W. C.

(1997). The multidimensional random

coefficients multinomial logit model. *Applied Psychological Measurement*, 11, 1–23.

<https://doi.org/10.1177/0146621697211001>

Amirian, S. M. R., Alavi, S. M., & Fidalgo, A. M.

(2014). Detecting gender DIF with an English proficiency test in EFL context. *Iranian Journal of Language Testing*, 4(2).

[http://ijlt.ir/content/ijlte\\_1\\_ov2\\_com/wp-content\\_138/uploads/2019/07/429-2014-4-2.pdf](http://ijlt.ir/content/ijlte_1_ov2_com/wp-content_138/uploads/2019/07/429-2014-4-2.pdf)

Barati, H., & Ahmadi, A. R. (2010). Gender-

based DIF across the subject area: A study of the Iranian National University Entrance Exam.

*The Journal of Teaching Language Skills*

(JTLS), 2(3), 1–22.

- NJ: Lawrence Erlbaum. <https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2330-8516.1986.tb00186.x>
- Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnosis modeling: Applying Wald test to investigate DIF for DINA model. *Journal of Educational Measurement, 11*, 98–125. <https://doi.org/10.1111/jedm.12036>
- Jang, E. E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign. [https://www.researchgate.net/profile/Eunice\\_Jang/publication/33746641](https://www.researchgate.net/profile/Eunice_Jang/publication/33746641)
- Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing, 66*(1), 31–73. <https://doi.org/10.1177/0265532208097336>
- Leighton, J., & Gierl, M. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press. <https://books.google.com/books>
- Li, F. M. (2008). A modified higher-order DINA model for detecting differential item functioning and differential attribute functioning. Unpublished doctoral dissertation, University of Georgia. <https://pdfs.semanticscholar.org/ebeb/8e863918509185f2e25f6031515a56c2b30c.pdf>
- Li, H. (2011). A cognitive diagnostic analysis of the MELAB reading test. *Spain Fellow, 9*, 17–46.
- validity: Construct representation versus nomothetic span. *Psychological Bulletin, 33*, 179–197. <https://psycnet.apa.org/record/1983-09487-001>
- Falmagne, J. C., & Doignon, J. P. (1988). A class of stochastic procedures for assessment of knowledge. *British Journal of Mathematical and Statistical Psychology, 11*, 1–23. <https://doi.org/10.1111/j.2044-8317>
- Finch, H. (2005). The MIMIC method as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement, 99*, 278–295. <https://doi.org/10.1177/0146621605275728>
- Gao, L., & Rogers, W. T. (2010). Use of tree based regression in the analyses of L2 reading test items. *Language Testing, 88*(2), 1–28. <https://doi.org/10.1177/0265532210364380>
- Haagenars, J., & McCutcheon, A. (2002). *Applied latent class analysis*. Cambridge: Cambridge University Press. <https://books.google.com/books>
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications. <https://books.google.com/books>
- Hartz, S. M. (2002). A bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign. <https://psycnet.apa.org/record/2002-95016-234>
- Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129–145). Hillsdale,

- Penfield, R. D., & Camilli, G. (2007). "Differential item functioning and item bias". In C.R. Rao & S. Sinharay (Vol. Eds.), *Handbook of statistics*, Vol. 26 (pp. 125 – 167), Elsevier. [https://doi.org/10.1016/S0169-7161\(06\)26005-X](https://doi.org/10.1016/S0169-7161(06)26005-X)
- Ranjbaran, F., & Alavi, S. M. (2017). Developing a reading comprehension test for cognitive diagnostic assessment: A RUM analysis. *Studies in Educational Evaluation*, 55, 167–179. <http://isiarticles.com/bundles/Article/pre/pdf/134341.pdf>
- Roeber, C. (2007). DIF in the Assessment of second language pragmatics. *Language Assessment Quarterly*, 4(2), 165–189. <https://doi.org/10.1080/15434300701375733>
- Rupp, A. A., & J., Templin. (2008). Unique characteristics of diagnostic classification models: a comprehensive review of the current state-of-the-art. *Meas Interdiscip Res Perspect*, 6, 219–262. <https://doi.org/10.1080/15366360802490866>
- Rupp, A. A, Templin, J, & R. A., Henson. (2010). *Diagnostic measurement: theory, methods, and applications*. Guilford, New York. <https://books.google.com/books>
- Shanmugam, S. K. S., & Lan, O. S. (2014). The validity of administering bilingual mathematics test among malasian bilingual students using Differential Item Function (DIF). *Asia Pacific Journal of Educators and Education*, 99, 1–18. [http://apjee.usm.my/APJEE\\_29\\_2014/Art%201\(1-18\).pdf](http://apjee.usm.my/APJEE_29_2014/Art%201(1-18).pdf)
- Shealy, R., & Stout, W. F. (1993a). An item response theory model for test bias. In P. W. Holland & H. Wainer (Eds.), <https://pdfs.semanticscholar.org/9916/0423205405cae7171c3f91e0215db2122947.pdf>
- Li, H. & Suen, H. K. (2013). Constructing and validating a Q-matrix for cognitive diagnostic analyses of a reading test, *Educational Assessment*, 88(1), 1–25. <https://doi.org/10.1080/10627197.2013.761522>
- Li, X. & Wang, W. C. (2015). Assessment of differential Iiem functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement*, 22, 28–54. <https://doi.org/10.1111/jedem.12061>
- Lim, Y. (2015). Cognitive diagnostic model comparisons. PhD Dissertation submitted to the Georgia Institute of Technology. <https://smartech.gatech.edu/bitstream/handle/1853/53513/LIMDISSERTATION-2015.pdf>
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hills-dale, NJ: Lawrence Erlbaum. [https://www.ets.org/research/policy\\_research\\_reports/publications/book/1980/jexj](https://www.ets.org/research/policy_research_reports/publications/book/1980/jexj)
- Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association*, 88, 690–700. <https://doi.org/10.1080/01621459.1963.10500879>
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Inst*, 22, 719–748. <https://doi.org/10.1093/jnci/22.4.719>
- Pae, T. I. (2004). Gender effect on reading comprehension with Korean EFL learners. *System*, 22(2), 265–281. <https://doi.org/10.1016/j.system.2003.09.009>

[https://arts.unimelb.edu.au/\\_data/assets/pdf\\_file/0010/1770679/Song\\_et\\_al.pdf](https://arts.unimelb.edu.au/_data/assets/pdf_file/0010/1770679/Song_et_al.pdf)

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 77, 361–370.  
<https://doi.org/10.1111/j.17453984.1990.tb00754.x>

Tatsuoka, K. K. (1983). Rule space: an approach for dealing with misconception based on item response theory. *Journal of Educational Measurement*, 00(4), 345–354.  
<https://doi.org/10.1111/j.17453984.1983.tb00212.x>

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Fredericksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Erlbaum.  
<https://psycnet.apa.org/record/1990-97343-018>

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale NJ: Erlbaum.  
<https://books.google.com/books>

Young, J. W., Morgan, R., Rybinski, P., Steinberg, J., & Wang, Y. (2013). Assessing the Test Information Function and Differential Item Functioning for the TOEFL Junior® Standard Test. *ETS Research Report Series*, 1, i-27.

*Differential item functioning* (pp. 197–329). Hillsdale, NJ: Lawrence Erlbaum.

<https://psycnet.apa.org/record/1993-97193-004>  
Shealy, R., & Stout, W. F. (1993b). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, 88, 159–194.  
<https://link.springer.com/article/10.1007/BF02294572>

Snow, R. E., & Lohman, D. F. (1989). *Implications of cognitive psychology for educational measurement*. American Council on Education.  
<https://psycnet.apa.org/record/1989-97348-007>

Song, X., Cheng, L., & Klinger, D. (2015). DIF investigations across groups of gender and academic background in a large-scale high-stakes language test.  
<https://www.ets.org/Media/Research/pdf/RR-13-17.pdf>

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, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.  
<https://s3.amazonaws.com/academia.edu.documents>

<https://s3.amazonaws.com/academia.edu.documents>

Zumbo, B. D. (2007). Three generations of DIF analysis: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4(2), 223–233.

<https://doi.org/10.1080/15434300701375832>

