

کاربرد مدل دو عاملی چندبعدي نظريه سوال-پاسخ جهت تحليل بعديت و کارکرد افتراقي سؤال، سؤالات مبتني بر متن

محمد احمدی ده قطب الدینی^۱، ابراهیم خدایی^{۲*}، ولی الله فرزاد^۳، علی مقدم زاد^۴، مسعود کبیری^۵

تاریخ دریافت: ۹۶/۰۳/۱۳

تاریخ پذیرش: ۹۶/۰۵/۱۵

چکیده

این مطالعه با هدف بررسی بعدیت و کارکرد افتراقی سؤال آزمون مبتنی بر متن پرلز ۲۰۱۱ ایران انجام گرفت. جهت تحلیل بعدیت از مدل‌های پاسخ مدرج شده و دو عاملی نظریه سوال-پاسخ با روش برآورد اطلاعات کامل پیشینه درست‌نمایی و جهت بررسی کارکرد افتراقی سؤال از مدل دو عاملی چندگروهی کای و همکاران (۲۰۱۱) استفاده شد. نتایج بررسی بعدیت نشان داد مدل دو عاملی نسبت به مدل پاسخ مدرج شده برازش بهتری با داده‌ها هم در کل نمونه ایرانی و هم در گروه پسران و دختران دارد. نتایج واریانس اثر متن نشان داد اثر عوامل ثانوی بر عملکرد دانش‌آموزان ایرانی در دو متن مرتبط با درک مطلب ادبی موجب بعدیت در متن‌های پرلز ایران شده است. نتایج نشان داد بین متوسط عملکرد دانش‌آموزان پسر و دختر در صفت نهفته کلی درک مطلب تفاوت معناداری وجود ندارد اما در سه متن ادبی و سه متن اطلاعاتی بین متوسط کارآمدی خواندن پسران و دختران تفاوت معناداری به نفع دختران وجود دارد. نتایج کارکرد افتراقی سؤالات بر مبنای مدل دو عاملی نیز نشان داد بیشتر سؤالات دارای کارکرد افتراقی یکپارچه و غیر یکپارچه هستند و پسران بیشتر در سؤالات چندگزینه‌ای و

۱. دانشجوی دکتری سنجش و اندازه‌گیری، دانشگاه تهران

۲. * دانشیار گروه روش‌ها و برنامه‌های آموزشی و درسی دانشگاه تهران khodaie@ut.ac.ir

۲. دانشیار گروه روان‌شناسی دانشگاه خوارزمی

۴. استادیار گروه روش‌ها و برنامه‌های آموزشی و درسی دانشگاه تهران

۵. کارشناس مرکز ملی مطالعات بین‌المللی پرلز و تیمز

دختران بیشتر در سؤالات پاسخ‌ساز عملکرد بهتری دارند. به طور کلی نتایج نشان داد در متن‌های پرلز ۲۰۱۱ ایران صفات مرتبط با دو متن درک مطلب ادبی در بین دانش‌آموزان پسر و دختر ایرانی به گونه متفاوتی ادراک می‌شوند و این دو متن در بین دختران وابستگی مکانی سؤال بیشتری نسبت به پسران دارند. همچنین نتایج حاکی از تفاوت عملکرد دانش‌آموزان پسر و دختر ایرانی در سؤالات آزمون شکل آمیخته پرلز است.

واژگان کلیدی: مدل پاسخ‌مدرج شده، مدل دو عاملی چندبعدي نظریه سوال-پاسخ، بعدیت، کارکرد افتراقی سؤال، آزمون‌های مبتنی بر متن

مقدمه

امروزه تحقیق مبتنی بر سواد آنقدر اهمیت پیدا کرده است که در طی چند دهه اخیر سنجش بین‌المللی کلان‌مقیاس سواد خواندن^۱ (پرلز) برای دانش‌آموزان پایه چهارم شکل گرفته است. این مطالعه تاکنون نه تنها نقش مهمی در ارزشیابی نظام‌های آموزشی جهان داشته است بلکه اطلاعات غنی و داده‌های تجربی برای توصیف پیامدهای آموزشی، ساختار کلی و جنبه‌های مهم نظام‌های آموزشی فراهم آورده است (لین‌کیث، چان، هاپفین‌بک و بایرد^۲، ۲۰۱۵). موضوعی که در ارتباط با اندازه‌گیری سواد خواندن در بین فرهنگ‌ها و زبان‌ها مختلف وجود دارد این است که اندازه‌گیری معنادار سواد نیازمند ابزارهایی است که به لحاظ فرهنگی و زبانی برای همه گروه‌های شرکت‌کننده هم‌ارز^۳ (معادل) باشند (کیرسچ، دی‌جانگ، لافونتاینه، مک‌کویین، میندیلوویتس و مانسیر^۴، ۲۰۰۲). هم‌ارزی آزمون‌های بین‌المللی باید از ابعاد مختلف همچون هم‌ارزی سازه^۵، اندازه‌گیری^۶ و تکنیک‌های جمع‌آوری داده‌ها^۷ مورد بررسی قرار گیرد. جهت تعیین

1. progress in international reading literacy study (PIRLS)
2. lenkeit, chan, hopfenbeck & baird
3. equivalent
4. kirsch, de jong, lafontaine, McQueen, mendelovits & monseur
5. construct equivalence
6. measurement equivalence
7. equivalence of data collection techniques

هم‌ارزی باید ابتداء سوگیری‌های سازه، سؤال و روش مورد بررسی قرار گیرد، فقدان سوگیری می‌تواند نشان دهنده هم‌ارزی باشد.

در سنجش‌های مبتنی بر متن همچون پرلز عوامل مختلفی می‌تواند منجر به سوگیری سازه یا بعدیت سؤالات آزمون شوند. اول اینکه در مطالعه پرلز همان‌طور که ریجمن^۱ (۲۰۱۱) بیان می‌کند سؤالات متن‌های خواندن در سطوح چندگانه‌ای خوشه‌بندی شده‌اند. مطالعه پرلز شامل بلوک‌هایی (آزمونچه‌ها^۲) از سؤال است. این بلوک‌های سؤال یا متن‌ها شامل حوزه‌های محتوایی هستند که به شکل متن‌های خواندن ارائه می‌شوند و با یک مجموعه سؤال چندگزینه‌ای و پاسخ‌ساز دنبال می‌شوند. سؤالات هر متن همگی موضوع واحدی را که در متن مطرح شده است می‌سنجند. ادراک متفاوت از این متن‌ها در بین گروه‌های مختلف می‌تواند بر بعدیت آزمون تأثیر بگذارد. سؤالات هر یک از این متن‌ها نیز در درون چهار فرایند اصلی خواندن یعنی تمرکز و بازیابی اطلاعات بیان شده به طور صریح، استنباط‌های مستقیم، تفسیر و تلفیق ایده‌ها و اطلاعات و بررسی و ارزیابی محتوا، زبان و عناصر متنی آشیانه کرده‌اند. همچنین هر یک از این فرایندهای خواندن در درون دو هدف اصلی خواندن یعنی خواندن جهت اکتساب و استفاده از اطلاعات و خواندن برای تجربه سواد آشیانه کرده‌اند. در نهایت همه ابعاد سواد خواندن در درون عامل اصلی درک مطلب آشیانه کرده‌اند. ریجمن (۲۰۱۱) مدل‌های سلسله‌مراتبی و مرتبه بالاتر متعددی را درباره داده‌های پرلز ۲۰۰۶ بکار برد و در نهایت با مدل لجستیک تک‌بعدي دو پارامتری^۳ به عنوان مدل پایه مقایسه کرد. نتایج نشان داد مدل سلسله‌مراتبی دو عاملی که در آن سؤالات در درون بلوک‌ها آشیانه کرده‌اند نسبت به دیگر مدل‌هایی که فرایندهای اصلی و اهداف خواندن در آن آشیانه کرده‌اند، برازش بهتری با داده‌ها دارد؛ بنابراین اثراتی از بعدیت در داده‌ها وجود دارد. بعدیت در داده‌ها، مفروضه مهم مدل‌های سنتی نظریه سوال-پاسخ یعنی تک‌بعدي بودن^۴ را به

1. rijmen
2. testlet
3. two-parameter logistic unidimensional item response theory model
4. uni-dimensionality

خطر می‌اندازد، در این حالت استفاده از مدل‌های چندبعدی نظریه سوال-پاسخ می‌تواند به فهم بهتر آنچه یک سؤال می‌سنجد، تعیین سطح کارآیی یک آزمودنی در هر صفت و اینکه با چه دقتی ترکیب‌های مختلف یک توانایی سنجش می‌شود کمک کند (سوکسوان، جان‌پینگ، این‌گودگراتوگ، گوی‌جاریرن پانیشگ^۱، ۲۰۱۲). آلپرکوز و دیمیرتاسلی^۲ (۲۰۱۲) بیان می‌کنند خطر تحلیل داده‌های چندبعدی به صورت تک‌بعدی این است که می‌تواند منجر به نمره خطای بیشتر در برآورد پارامترهای توانایی و سؤالات از یک سو و کاهش دقت اندازه‌گیری و برازش داده‌ها با مدل از سوی دیگر شود در نتیجه نتایج به دست آمده دارای سوگیری خواهند بود.

دوم اینکه همان‌طور که فوکوهارا^۳ (۲۰۰۹) بیان می‌کند در آزمون‌های مبتنی بر متن همچون پرلز سؤالات درون یک متن ممکن است به طور کلی از یکدیگر مستقل نباشند زیرا سطح فهم متن خواندن ممکن به وسیله دانش و علاقه آزمودنی‌ها درباره محتوای متن خواندن تحت تأثیر قرار گیرد. این امر منجر به تخطی از مفروضه مهم نظریه‌های اندازه‌گیری یعنی استقلال مکانی سؤالات^۴ می‌شود. تائو^۵ (۲۰۰۸) بیان می‌کند این نوع وابستگی یک وابستگی ترکیب^۶ است و در موقعیتی به وجود می‌آید که سؤالات وابسته به یک محرک یکسان (مثل سؤالات مربوط به یک متن خواندن در آزمون درک مطلب) باشند. فوکوهارا (۲۰۰۹) بیان می‌کند تخطی از این مفروضه پیامدهایی نامطلوبی دارد. از چشم انداز نظریه کلاسیک آزمون به واسطه استفاده از متن‌ها، خطای استاندارد اندازه‌گیری کم برآورد و در نتیجه پایایی بیش برآورد می‌شود و از چشم انداز نظریه سوال-پاسخ، تابع اطلاعات بیش برآورد می‌شود و این به معنای این است که خطای استاندارد

1. suksuan, junpeng, ngudgratoke, and guayjarernpanishk
2. Alperkose & Demirtasli
3. Fukuhara
4. local item independence
5. Tao
6. combination dependence

برآورد توانایی کم برآورد شده است. همچنین ممکن است منجر به سوگیری در قدرت تمیز سؤالات شود.

سوم اینکه در مطالعه پرلز جهت سنجش سواد خواندن از متن‌های دارای سؤالات با شکل آمیخته^۱ استفاده می‌شود یعنی سؤالات متن‌ها هم به شکل چندگزینه‌ای و هم پاسخ‌ساز است. تراپ^۲ (۱۹۹۳)، به نقل از مونتگومری^۳، (۲۰۱۴)، بیان می‌کند اثر شکل^۴ در نتیجه استفاده از فرایندهای ذهنی متفاوت جهت پردازش سؤالات با شکل متفاوت توسط آزمودنی به وجود می‌آید. کیم و کولن^۵ (۲۰۰۶)، به نقل از مونتگومری، (۲۰۱۴) نیز بیان می‌کنند زمانی آزمون شکل آمیخته داشته باشد نسبت به زمانی که شکل ساده (یک نوع سؤال) دارد سطحی از پیچیدگی بیش از آزمون‌های با شکل ساده اعمال می‌شود. زمانی که اثر شکل آزمون روی می‌دهد سؤالات چندگزینه‌ای و پاسخ‌ساز، توانایی‌های متفاوتی را اندازه‌گیری می‌کنند و موجب حضور چند بعدیتی در نمره کل آزمون می‌شوند.

چهارم اینکه همان‌طور که راثوچ و هارتیک^۶ (۲۰۱۰) بیان می‌کنند متغیرهای گروه‌بندی شده‌ای همچون جنسیت، زبان خواندن در خانه، ناحیه‌ای که مدرسه در آن قرار دارد و وضعیت اقتصادی-اجتماعی خانواده در ارتباط با خواندن شناسایی شده‌اند که می‌توانند بر عملکرد دانش‌آموزان گروه‌های مختلف در متن‌های خواندن تأثیر بگذارند. در گزارش پینزا^{۲۰۰۶} بیان شده است که خواندن حوزه‌ای است که شکاف جنسیتی در آن زیاد است و دختران نمرات بالاتری نسبت به پسران کسب می‌کنند، همچنین آزمون‌دهندگان که به زبانی غیر از زبان آزمون

-
1. mixed item format
 2. Traub
 3. Montgomery
 4. format effect
 5. Kim & Kolen
 6. Rauch & Hartig

در خانه صحبت می‌کنند انتظار می‌رود عملکرد پایین‌تری در سنجش خواندن نسبت به آزمون دهندگانی که از زبان آزمون برای برقراری ارتباط در خانه استفاده می‌کنند داشته باشند. جهت بررسی اثر این متغیرهای گروه‌بندی شده باید سوگیری سؤال یا آنچه به کارکرد افتراقی سؤال^۱ معروف است انجام گیرد (فوکوهارا، ۲۰۰۹). سوگیری سؤال اشاره به غیرمتعارف بودن یک ابزار در سطح سؤال می‌کند (ون‌دِ ویجور و لئونگ^۲، ۱۹۹۷) و می‌تواند هم‌ارزی ابزارها را به خطر بیندازد. در مطالعات پیمایشی بین کشورها، هم‌ارزی سؤالات پرسشنامه یک مجادله دائم بین روایی و پایایی سؤالات است (جانسون^۳، ۱۹۹۸). بلوم، گلدشتاین و گورین-پیس^۴ (۲۰۰۱)، به نقل از گلدشتاین، (۲۰۰۸) بیان می‌کنند سؤالاتی که به طور متفاوتی در بعضی کشورها عمل می‌کنند به نظر سوگیری فرهنگی دارند. این سؤالات ممکن در برگیرنده اطلاعات مهم باشند و منعکس کننده تفاوت‌های جالب توجه در پاسخ‌های بین کشورها باشد. دِوینس، بیزر، دیون، پلیتیر و دواردز^۵ (۱۹۹۷) بیان می‌کنند در این میان نظریه سوال-پاسخ می‌تواند برای شناسایی سؤالاتی از یک پرسشنامه که به روش مشابه یا نامشابه بین فرهنگ‌ها عمل می‌کنند به کار برده شود. این تکنیک توسط روان‌شناسان برای شناسایی سؤالات آزمون‌ی که سازه نهفته زیربنایی مورد اندازه‌گیری را منعکس نمی‌کنند به کار برده می‌شود. در این نظریه این کار با مقایسه منحنی ویژگی سؤال^۶ (احتمال شرطی پاسخ به سؤال در سطوح مختلف متغیر نهفته‌ای که توسط مقیاس اندازه‌گیری ارائه می‌شود) در بین گروه‌ها انجام می‌گیرد. شباهت منحنی‌های ویژگی سؤال بین گروه‌های فرهنگی به عنوان شاهدهی از رفتار یکسان و بنابراین هم‌ارزی تفسیر می‌شود، اما مشکل اصلی مدل‌های سنتی این است که مفروضه تک‌بعدی بودن سؤالات آزمون باید برقرار باشد.

1. differential item functioning(DIF)
2. Van de Vijver & Leung
3. Johnson
4. Blum, Goldstein & Guerin-pace
5. Devins, Beiser, Dion, Pelletier & Edwards
6. item characteristic curve

در طی سال‌های اخیر تلاش‌هایی برای تدوین مدل‌های ترکیبی با ترکیب دو مکتب فکری نظریه دو عاملی اسپیرمن^۱ و ساختار ساده^۲ ثرستون^۳ از عوامل چندگانه توانایی‌های ذهنی نخستین^۴ با نظریه سوال-پاسخ صورت گرفته که منجر به تدوین مدل‌های مختلف چندبعدی در حوزه نظریه سوال-پاسخ شده است (امدی دیسا،^۵ ۲۰۱۲). این مدل‌ها ترکیبی از مدل‌های لجستیک^۶ برای سؤالات دوارزشی^۷ مثل مدل دو و سه پارامتری و مدل‌های سؤالات نمره‌گذاری شده به صورت چند ارزشی^۸ مثل مدل پاسخ مدرج هستند. از جمله مدل‌هایی چندبعدی که در سال‌های اخیر در سنجش‌های آموزشی کاربرد زیادی پیدا کرده است مدل دو عاملی چندبعدی نظریه سوال-پاسخ^۹ است. در مدل دو عاملی چندبعدی نظریه سوال-پاسخ یک بعد کلی^{۱۰} و چند بعد اختصاصی^{۱۱} (بسته به عامل‌های مورد بررسی مثل شکل سؤالات، متن‌ها، فرایندهای پردازش و...) وجود دارد (کای، یانگ و هانسن^{۱۲}، ۲۰۱۱). ابعاد اختصاصی یا متن‌ها جهت تبیین وابستگی‌های اضافی اعمال شده همچون کاربرد متن، مورد استفاده قرار می‌گیرند؛ بنابراین در این مدل یک پارامتر اضافی برای اثر متن^{۱۳} به مدل سنتی نظریه سوال-پاسخ اضافه می‌شود و به عنوان یک متغیر تصادفی در نظر گرفته می‌شود (فوکوهارا،^{۲۰۰۹}). وابستگی مکانی ایجاد شده بین سؤالات به دلیل اثر متن، یک بعد اضافی به آزمون اضافه می‌کند که باید در گسترش سازه هدف مورد نظر

-
1. Spearman's two-factor theory
 2. simple structure
 3. Thurston
 4. theory of primary mental abilities
 5. Md Desa
 6. logistic models
 7. binary items
 8. polytomous items
 9. bi-factor multidimensional IRT model
 10. general dimension
 11. specific dimension
 12. cai, yang and hansen
 13. testlet effect

قرار گیرد (زینسکی، هامبلتون، سیرسی^۱، ۲۰۰۱). در این مدل جهت تحلیل سؤالات دوازده‌گانه از تابع پیوند لجیت^۲ و برای سؤالات چند ارزشی از تابع لجیت تراکمی^۳ استفاده می‌شود. در این روش با بررسی تفاوت لگاریتم نسبت درست‌نمایی^۴ مدل دو عاملی نظریه سوال-پاسخ با مدل تک بعدی دو پارامتری برای سؤالات دو ارزشی و مدل پاسخ مدرج شده برای سؤالات چندارزشی می‌توان به اثر متن‌ها بر ایجاد بعدیت در آزمون پی برد (کای و همکاران، ۲۰۱۱).

همچنین در مدل‌های سنتی نظریه سوال-پاسخ برای آشکارسازی کارکرد افتراقی سؤال، وابستگی مکانی ایجاد شده به واسطه متن‌ها در نظر گرفته نمی‌شود (فوکوهارا و کاماتا^۵، ۲۰۱۱). فوکوهارا (۲۰۰۹) بیان می‌کند در چند سال اخیر مدل‌های مختلفی برای بررسی کارکرد افتراقی سؤالات مبتنی بر متن شکل گرفته است. فوکوهارا (۲۰۰۹)، بیان می‌کند از جمله این مدل‌ها می‌توان به مدل پاسخ متن راش^۶ و آنکگ و ویلسون^۷ (۲۰۰۵)، مدل دو پارامتری پاسخ متن^۸ و آنکگ، بردلاو، واینر و مولر^۹ (۲۰۰۸)، مدل چندگروهی متن^{۱۰} لای، بولت و فو^{۱۱} (۲۰۰۶)، مدل چندگروهی دو عاملی^{۱۲} کای و همکاران (۲۰۱۱) و مدل دو عاملی فوکوهارا و کاماتا (۲۰۱۱) اشاره کرد. مشکل مدل پاسخ متن راش، تک پارامتری بودن و نداشتن پارامتر اثر^{۱۳} است. پارامتر اثر بیان‌کننده عملکرد افتراقی در سؤال بین دو گروه است که از تفاوت در متوسط توانایی‌های دو گروه به وجود می‌آید و منعکس‌کننده این است چگونه دو گروه در سطوح توانایی‌هایشان بر

1. zenisky, hambleton and sireci
2. logit link function
3. cumulative logit link function
4. log likelihood ratio
5. kamata
6. rasch testlet response model
7. wang and wilson
8. two parameter testlet response model
9. wang, bradlow, wainer and muller
10. multiple-group testlet model
11. li, bolt and fu
12. multiple-group bifactor model
13. impact effect

حسب عملکرد در سؤال با یکدیگر متفاوت هستند، در حالی که کارکرد افتراقی سؤال بیان می‌کند چگونه دو گروه به طور متفاوتی در سؤال در سطح یکسانی از توانایی عمل می‌کنند (فوکوهارا، ۲۰۰۹)، همچنین مدل‌های پاسخ متن‌راش و پاسخ متن دو پارامتری اجازه هیچگونه تفاوتی در توزیع‌های متغیرهای نهفته بین گروه‌ها را نمی‌دهند. در مدل فوکوهارا و کاماتا (۲۰۱۱) نیز فقط به میانگین بعد کلی اجازه داده می‌شود در بین گروه‌ها متفاوت باشد. همچنین فوکوهارا (۲۰۰۹) بیان می‌کند در کارکرد افتراقی سؤال یک ملاک همتا بودن برای کنترل تفاوت در توانایی بین دو گروه نیاز است. در رویکردهای سنتی نظریه سؤال-پاسخ نمره کل آزمون اغلب به عنوان ملاک همتا بودن^۱ به کار برده می‌شود اما در رویکردهای جدید کارکرد افتراقی مبتنی بر متن از سؤالات لنگر برای برآورد ملاک همتا بودن توانایی گروه‌ها استفاده می‌شود. سؤالات لنگر، سؤالات فاقد کارکرد افتراقی سؤال هستند.

مطالعاتی که از رویکردهای جدید حوزه نظریه سوال-پاسخ همچون مدل دو عاملی برای بررسی بعدیت و کارکرد افتراقی سؤال در آزمون مبتنی بر متن پرلز استفاده کرده باشد به جز مطالعه ریچمن (۲۰۱۱) که فقط بعدیت را در تمام داده‌های کشورهای شرکت کننده مطالعه پرلز ۲۰۰۶ بررسی کرده است، مطالعه دیگری در ارتباط با آزمون‌های مبتنی بر متن مشاهده نشد. در این مطالعه نیز سؤالات به صورت دو ارزشی تحلیل شده‌اند و به شکل آمیخته سؤالات توجه نشده است اما نتایج وجود اثراتی از تأثیر متن‌ها بر وابستگی مکانی مثبت و منفی بین سؤالات متن‌ها را نشان داد. از سوی دیگر مطالعاتی که داده‌های پرلز را بر حسب مدل دو عاملی از جنبه‌های هم‌ارزی سازه و سؤال و با توجه به عامل‌هایی همچون جنسیت، زبان، وضعیت اجتماعی-اقتصادی و ... مورد بررسی قرار داده باشند نیز مشاهده نشد. با توجه به تأثیر جنسیت به خصوص در آزمون‌های خواندن بر عملکرد شرکت کنندگان یک آزمون و فقدان مطالعاتی که از مدل دو عاملی برای بررسی اثر جنسیت بر بعدیت و کارکرد افتراقی سؤال متن‌های مطالعه پرلز استفاده کرده

1. matching criterion

باشند این مطالعه با هدف کاربرد مدل دو عاملی نظریه سوال-پاسخ جهت بررسی بعدیت و کارکرد افتراقی سؤالات متن‌های مطالعه پرلز ۲۰۱۱ ایران با توجه به جنسیت دانش‌آموزان انجام گرفت. به این منظور سؤالات زیر مطرح گردید:

آیا مدل دو عاملی چندبعدی نظریه سوال-پاسخ نسبت به مدل تک‌بعدی پاسخ مدرج شده برازندگی بهتری با سؤالات متن‌های خواندن مطالعه پرلز ۲۰۱۱ ایران دارد؟ (بعدیت) کدام متن‌های خواندن مطالعه پرلز ۲۰۱۱ ایران موجب انحراف از تک‌بعدی بودن داده‌ها شده است؟ (اثر متن)

آیا سؤالات متن‌های پرلز ۲۰۱۱ در بین دانش‌آموزان پسر و دختر ایرانی دارای پارامترهای سؤال (ضریب تمیز، آستانه دشواری) متفاوتی هستند؟

روش

این مطالعه یک مطالعه توصیفی است که بر روی داده‌های پیمایش بین‌المللی پرلز ۲۰۱۱ ایران انجام گرفت. به این منظور از بین ۴۹ کشور شرکت‌کننده در پرلز ۲۰۱۱، کشور ایران انتخاب و با توجه به جنسیت دانش‌آموزان بعدیت و کارکرد افتراقی سؤالات متن‌های مطالعه پرلز ۲۰۱۱ مورد تحلیل قرار گرفت. تعداد کل دانش‌آموزان ایرانی پایه چهارم ابتدایی شرکت‌کننده در مطالعه پرلز ۲۰۱۱ برابر ۵۷۵۸ نفر است که از این تعداد ۲۸۵۹ نفر (۴۹/۷ درصد) دانش‌آموز دختر و ۲۸۹۹ نفر (۵۰/۳ درصد) دانش‌آموز پسر بودند.

ابزار اندازه‌گیری پژوهش حاضر از قرار زیر است:

آزمون سواد خواندن مطالعه پرلز دارای ۱۰ متن است که در ۱۳ دفترچه قرار گرفته‌اند. در دفترچه ۱ تا ۳ هر دو متن ادبی و در دفترچه‌های ۵ تا ۷ هر دو متن اطلاعاتی و در دفترچه‌های ۴ و ۸ تا ۱۳ یک متن ادبی و یک متن اطلاعاتی هستند. در این ۱۳ دفترچه در مجموع ۱۰ متن با ۱۴۶ سؤال مطرح شده است که ۵ متن ادبی و ۵ متن اطلاعاتی هستند. در هر متن نیز سؤالات به

شکل چندگزینه‌ای و پاسخ‌ساز هستند. از ۱۴۶ سؤال ۷۵ سؤال (۵۱/۳۷ درصد) ادبی و ۷۱ سؤال (۴۱/۶۳ درصد) اطلاعاتی هستند. همچنین از این تعداد سؤال ۶۲ سؤال (۴۲/۴۷ درصد) چندگزینه‌ای و ۸۴ سؤال (۵۷/۵۳ درصد) پاسخ‌ساز هستند. نام این ده متن به ترتیب عبارت است از گل‌های روی پشت بام (۱۳ سؤال)، پرواز کن عقاب، پرواز کن (۱۲ سؤال)، کهربای درخشان (۱۴ سؤال)، گلدان خالی (۲۰ سؤال)، کلوچه دشمن (۱۲ سؤال)، لئوناردو داوینچی (۱۲ سؤال)، لذت پیاده روی روزانه (۱۲ سؤال)، کوسه‌ها (۱۶ سؤال)، در جستجوی عسل (۱۶ سؤال) و راز دندان غول (۱۹ سؤال).

جهت تعیین بعدیت داده‌ها همان‌طور که دی‌مارس^۱ (۲۰۱۳) مطرح می‌کند باید شاخص‌های برازندگی دو مدل تک‌بعدی دو پارامتری لجستیک برای داده‌های دو ارزشی و مدل پاسخ مدرج شده برای داده‌های چند ارزشی با شاخص‌های مدل دو عاملی نظریه سوال-پاسخ مقایسه شود. از آنجایی که سؤالات متن‌های پرلز به شکل آمیخته هستند از مدل تک‌بعدی پاسخ مدرج شده جهت بررسی تک‌بعدی بودن داده‌ها و از مدل دو عاملی نظریه سوال-پاسخ جهت تعیین اثر متن‌ها بر انحراف از تک‌بعدی بودن استفاده شد. جهت مقایسه این مدل‌ها از آزمون لگاریتم نسبت درست‌نمایی^۲، ملاک اطلاعات آکائیک^۳ (AIC) و ملاک اطلاعات بیزی^۴ (BIC) استفاده می‌شود. دوآن، هاردل و جینتل^۵ (۲۰۱۲) بیان می‌کنند ساده‌ترین شیوه، مقایسه تابع لگاریتم درست‌نمایی مدل پایه و مدل آشیانه‌ای است اما از آنجایی که مدل‌ها دارای تعداد پارامترهای یکسان نیستند و از آنجایی که مدل‌ها با تعداد پارامتر بیشتر، بیشتر اوقات برازش بهتری با داده‌ها نشان می‌دهند باید از شاخص‌های ملاک اطلاعات آکائیک و ملاک اطلاعات بیزی استفاده شود

1. DeMars
2. log likelihood ratio test
3. akaike information Criterion (AIC)
4. bayesian information Criterion (BIC)
5. duan, hardle and gentle

که جریمه‌های متنوعی برای تعداد پارامترهای برآورده شده در نظر می‌گیرند. مدلی که ملاک اطلاعات آکائیک و بیزی کوچک‌تر داشته باشد، برازندگی بهتری با داده‌ها دارد.

جهت تعیین اینکه کدام متن‌ها موجب انحراف از تک‌بعدی بودن شده‌اند یا اینکه کدام متن‌ها دارای وابستگی مکانی سؤال هستند و جنسیت دانش‌آموزان چند درصد از واریانس این وابستگی را تبیین می‌کند از مدل دو سطحی دو عاملی استفاده شد. در مدل دو سطحی دو عاملی دو نوع واریانس تصادفی درون گروهی (واریانس در مدل بدون پیش‌بینی کننده) و بین گروهی (واریانس در مدل با پیش‌بینی کننده) برای هر یک از متن‌ها وجود دارد. واریانس درون گروهی برای تعیین اثر متن یا وجود وابستگی مکانی بین سؤالات متن و واریانس بین گروهی برای کمک به تعیین میزان تأثیر متغیر بین گروهی (در این مطالعه جنسیت) بر وجود وابستگی مکانی بین سؤالات متن‌ها به کار برده می‌شود. اگر واریانس تصادفی اثر متن صفر باشد، استقلال مکانی سؤال در بین سؤالات متن وجود دارد اما اگر صفر نباشد، وابستگی مکانی سؤال بین سؤالات یک متن را نشان می‌دهد. طبق مطالعات شبیه‌سازی واریانس تصادفی اثر متن کوچکتر از ۰/۲۵ نشان دهنده وابستگی مکانی سؤال ناچیز در بین سؤالات متن و بالاتر از ۰/۲۵ نشان دهنده وابستگی مکانی قابل ملاحظه بین سؤالات متن است (ژانگ، سین و کندی، ۲۰۱۰؛ به نقل از روند، ۲۰۱۵). بر طبق مطالعات تجربی نیز واریانس تصادفی اثر متن بین ۰/۵ تا ۲ نشان دهنده وابستگی مکانی قابل ملاحظه بین سؤالات متن است (اکس و آنک، برادلاو و واینر، ۲۰۰۲ و بی. ژانگ، ۲۰۱۰؛ به نقل از روند، ۲۰۱۵). به منظور تعیین اینکه تا چه اندازه وابستگی مکانی ایجاد شده در متن حاصل عامل پیش‌بینی کننده (جنسیت دانش‌آموزان) است، واریانس اثر متن در مدل بدون پیش‌بینی کننده با واریانس اثر متن در مدل با پیش‌بینی کننده (عامل زبان) بر حسب روش

1. zhang, shen and cannady
2. x. wang, bradlow and wainer
3. b. zhang

بریگ و راثو دین بوش^۱ (۱۹۹۲، به نقل از روند، ۲۰۱۵) مورد مقایسه قرار گرفت. این روش به شرح زیر است:

(واریانس اثر متن در مدل بدون پیش بین - واریانس اثر متن در مدل با پیش بین)

واریانس اثر متن در مدل بدون پیش بین

جهت تحلیل کارکرد افتراقی سؤال نیز از رویکرد معرفی شده توسط کای و همکاران (۲۰۱۱) استفاده شد. در این روش برای انجام تحلیل کارکرد افتراقی سؤال از رویکرد نسبت درست‌نمایی نظریه سوال-پاسخ برای کارکرد افتراقی سؤال^۲ (IRTLRDIF) استفاده می‌شود. این رویکرد یک روش مبتنی بر نظریه سوال-پاسخ است. در این روش در مرحله اول باید مقیاس متغیرهای نهفته پیوند داده شود. به این منظور فرض می‌شود در مدل پایه یا مدل فاقد کارکرد افتراقی سؤال همه پارامترهای سؤالات در بین گروه‌ها برابر هستند. در مدل مقایسه برای مطالعه کارکرد افتراقی سؤال محدودیت‌های برابری بین گروهی خوشه‌های سؤال نادیده گرفته می‌شود. جهت تحلیل کارکرد افتراقی سؤالات هر متن، سؤالات آن متن به عنوان داوطلب جهت بررسی کارکرد افتراقی سؤال قرار داده می‌شود و سؤالات باقی‌مانده در متن‌های دیگر به عنوان لنگر به کار برده می‌شوند تا مقیاس متغیرهای نهفته به یکدیگر پیوند داده شوند. برای بررسی کارکرد افتراقی سؤال، عرض از مبدا سؤال و شیب‌ها روی عامل کلی در هر گروه برای خوشه‌های سؤال آزادانه برآورد می‌شود و از سوی دیگر شیب‌های سؤال روی عامل خاص برای خوشه سؤال هم در درون گروه و هم بین گروه به طور همزمان برابر قرار داده می‌شوند و واریانس‌های عامل خاص برای همه گروه‌ها به جزء گروه مرجع که فرض می‌شود واریانس آن برابر یک است آزادانه برآورد می‌شود (کای و همکاران، ۲۰۱۱). جهت برآورد پارامترها نیز از روش برآورد اطلاعات کامل بیشینه درست‌نمایی برای داده‌های ناکامل استفاده شد. چن و فین برگ^۳ (۱۹۷۴)، به نقل از لینگ

1. bryk and raubenbush
2. item-reponse theory likelihood ratio differential item functioning (IRTLRDIF)
3. chen and fienberg

پینگ و اسلام^۱، ۲۰۰۸) روش بیشینه درست‌نمایی برای جداول توافقی ناکامل هنگامی که ردیف‌ها و ستون‌های داده‌ها یا هر دو دارای مقادیر گم شده هستند به کار برده می‌شود. به منظور استفاده از این روش از الگوریتم انتظار-بیشینه‌سازی استفاده شد. در الگوریتم انتظار-بیشینه‌سازی در هر تکرار دو مرحله انتظار و بیشینه‌سازی وجود دارد و هدف بهبود همگرایی بین دو مرحله است. این الگوریتم برای تحلیل داده‌های ناکامل طبقه‌ای نیز به کار برده می‌شود.

جهت برآورد پایایی کل آزمون و آزمونچه‌ها نیز از پایایی تجربی یا حاشیه‌ای^۲ استفاده شد. از آنجایی که در نظریه سوال-پاسخ Θ مشخص نیست و خطای استاندارد بین آزمودنی یکسان نیست پایایی حاشیه‌ای می‌تواند به صورت $S^2_{\tau} = 1 - (S^2_e / S^2_{\tau})$ برآورد شود که S_e متوسط خطای استاندارد و S_{τ} انحراف استاندارد برآورد شده توزیع جامعه (مجموع واریانس نمرات پسینی مورد انتظار (EAP) و واریانس خطا) است. نتایج پایایی در جدول ۱ آورده شده است. نتایج نشان داد که پایایی آزمونچه‌ها در کل نمونه ایرانی نسبت به زیر گروه‌های دختران و پسران از پایایی مطلوب-تری برخوردار هستند.

جدول ۱. نتایج پایایی تجربی کل آزمون و آزمونچه‌ها
در نمونه کلی و گروه دختران و پسران ایرنای مطالعه پرلز ۲۰۱۱

گروه	کل آزمون	آزمونچه									
		اول	دوم	سوم	چهارم	پنجم	ششم	هفتم	هشتم	نهم	دهم
کل	۰/۸۴	۰/۸۷۷	۰/۸۱۳	۰/۸۵۸	۰/۸۱۲	۰/۸۳۲	۰/۸۵۸	۰/۸۳۳	۰/۸۳۴	۰/۸۴۱	۰/۸۴۴
پسر	۰/۶۱۹	۰/۶۳۱	۰/۵۱۹	۰/۳۹۷	۰/۶۸۴	۰/۵۸۷	۰/۶۳۴	۰/۵۷۶	۰/۶۷۵	۰/۶۷۲	۰/۶۱۱
دختر	۰/۵۹۴	۰/۶۰۶	۰/۶۰۹	۰/۵۱۸	۰/۷۱۲	۰/۵۲۸	۰/۵۶۲	۰/۳۲۲	۰/۵۸۹	۰/۵۶۶	۰/۶۲۶

1. ling ping and islam
2. emperical or marginal reliability

یافته‌ها

الف: تعیین بعدیت. جدول ۲ نتایج بررسی بعدیت آزمون سواد خواندن پرلز ۲۰۱۱ ایران در کل نمونه و در گروه دانش آموزان پسر و دختر بر مبنای مدل‌های پاسخ مدرج شده و دو عاملی چندبعدی نظریه سوال-پاسخ را نشان می‌دهد. نتایج بررسی دو مدل در کل نمونه نشان داد شاخص‌های برازندگی در مدل پاسخ مدرج شده ($BIC=283627/182$, $AIC=280338/492$), $p < 0/01$ ($-2LL_{(660)} = -139509/246$) و مدل دو عاملی ($BIC=277215/832$), $p < 0/01$ ($-2LL_{(806)} = -134865/88$, $AIC=271343/761$) نشان دهنده برازندگی بهتر مدل دو عاملی نسبت به مدل پاسخ مدرج شده با داده‌ها است؛ زیرا لگاریتم نسبت درست‌نمایی، ملاک اطلاعات آکائیک و ملاک اطلاعات بیزی، مدل دو عاملی نسبت به مدل پاسخ مدرج شده کوچکتر هستند همچنین نتایج مقایسه تفاوت لگاریتم نسبت درست‌نمایی دو مدل نیز معنادار است

($p < 0/01$, $-2LL_{(145)} = -4643/366$)، در نتیجه وجود اثراتی از اثر متن در داده‌های پرلز ایران وجود دارد. نتایج شاخص‌های برازندگی مدل پاسخ مدرج شده ($BIC=130634/525$), $p < 0/01$ ($-2LL_{(519)} = -63552/014$, $AIC=128142/029$) و مدل دو عاملی ($BIC=130414/229$), $p < 0/01$ ($-2LL_{(665)} = -62352/651$) در گروه پسران و نتایج شاخص‌های برازندگی مدل پاسخ مدرج شده ($BIC=124304/347$), $p < 0/01$ ($-2LL_{(523)} = -60493/024$, $AIC=122032/048$) و مدل دو عاملی ($BIC=124278/748$), $p < 0/01$ ($-2LL_{(669)} = -59283/051$, $AIC=119904/101$) در گروه دختران نشان داد مدل دو عاملی هم بر پایه شاخص لگاریتم نسبت درست‌نمایی و هم بر پایه ملاک اطلاعات آکائیک و بیزی برازندگی بهتر با داده‌ها هم در گروه پسران و دختران دارد. تفاوت لگاریتم نسبت درست‌نمایی دو مدل در گروه پسران ($p < 0/01$, $-2LL = 1199/363$)

$\Delta(146) =$ و دختران ($p < 0/01$ ، $-1209/973 = \Delta - 2LL(146)$) نیز معنادار است در نتیجه اثراتی از اثر متن در داده‌های پرلز دختران و پسران ایرانی نیز وجود دارد.

جدول ۲. شاخص‌های تعیین بعدیت در آزمون سواد خواندن ایران

در حالت کلی و در گروه پسران و دختران

شاخص	مدل	کل ایران	پسر	دختر
۲- لگاریتم نسبت درست‌نمایی (-) (2LL)	دو پارامتری	-۱۳۹۵۰۹/۲۴۶	-۶۳۵۵۲/۰۱۴	-۶۰۴۹۳/۰۲۴
	دو عاملی	-۱۳۴۸۶۵/۸۸	-۶۲۳۵۲/۶۵۱	-۵۹۲۸۳/۰۵۱
درجه آزادی	دو پارامتری	۶۶۰	۵۱۹	۵۲۳
	دو عاملی	۸۰۶	۶۶۵	۶۶۹
تفاوت بین مقادیر خن دو مدل	-	-۴۶۴۳/۳۶۶	-۱۱۹۹/۳۶۳	-۱۲۰۹/۹۷۳
تفاوت بین درجه آزادی دو مدل	-	۱۴۶	۱۴۶	۱۴۶
سطح معنی‌داری	-	$P < 0/01$	$P < 0/01$	$P < 0/01$
ملاک اطلاعات آکائیک	دو پارامتری	۲۸۰۳۳۸/۴۹۲	۱۲۸۱۴۲/۰۲۹	۱۲۲۰۲۲/۰۴۸
	دو عاملی	۲۷۱۳۴۳/۷۶۱	۱۲۶۰۳۵/۳۰۲	۱۱۹۹۰۴/۱۰۱
ملاک اطلاعات بیزی	دو پارامتری	۲۸۳۶۲۷/۱۸۲	۱۳۰۶۳۴/۵۲۵	۱۲۴۳۰۴/۳۴۷
	دو عاملی	۲۷۷۲۱۵/۸۳۲	۱۳۰۴۱۴/۲۲۹	۱۲۴۲۶۷/۷۴۸

پ: تعیین اثر متن (وابستگی مکانی سؤال) بر بعدیت. جدول ۳ نتایج تحلیل بارهای عاملی (شیب یا ضریب تمیز) و ضریب تبیین سؤالات مطالعه پرلز ۲۰۱۱ را در کل نمونه ایرانی و دانش‌آموزان پسر و دختر نشان می‌دهد (در این مقاله به علت تعداد زیاد جداول فقط جدول مقایسه متن اول در بین گروه‌های مورد مطالعه آورده شده است). نتایج نشان داد در کل نمونه ایرانی و نمونه پسران و دختران ایرانی در مدل پاسخ مدرج شده بارهای عاملی روی صفت نهفته مثبت هستند اما اکثر این بارهای عاملی کوچک (زیر ۰/۸) هستند و مقدار واریانس تبیین شده سؤالات توسط صفت نهفته زیربنایی کم است این مطلب می‌تواند حاکی از وجود عوامل نهفته ثانویه باشد که بر سؤالات آزمون اثر می‌گذارد. جهت بررسی اثر احتمالی عوامل اضافی (متن‌ها) تاثیرگذار بر

سؤالات از مدل دو عاملی نظریه سوال-پاسخ استفاده شد. نتایج نشان داد که درصد بیشتری از واریانس سؤالات توسط عامل کلی و عوامل خاص (درک مطلب ادبی و اطلاعاتی) تبیین می‌شود و بارهای عاملی اکثر سؤالات در کل نمونه ایرانی و نمونه پسران و دختران روی عامل خاص بزرگتر از صفت کلی هستند که می‌تواند حاکی از وجود اثرات عوامل ثانویه بر سؤالات و به عبارت دیگر انحراف از تک‌بعدی بودن متن‌های خواندن پرلز در کل نمونه ایرانی و نمونه پسران و دختران باشد. نتایج مقدار واریانس تبیین شده سؤالات در مدل تک‌بعدی پاسخ مدرج شده و مدل دو عاملی نیز حاکی از افزایش دقت برآورد سؤالات در مدل دو عاملی توسط عامل کلی و عوامل خاص نسبت به مدل تک‌بعدی پاسخ مدرج شده است.

نتایج جدول ۴ نیز نشان می‌دهد در مطالعه پرلز ۲۰۱۱ ایران دو متن مرتبط با سنجش درک مطلب ادبی (صفات نهفته خاص اول و دوم مرتبط با متن‌های گل‌های روی پشت بام و پرواز کن عقاب پرواز کن) دارای واریانس تصادفی اثر متن ۰/۴۰۷ و ۰/۳۵۷ هستند که بر طبق نتایج مطالعات شبیه‌سازی نشان دهنده وابستگی مکانی سؤال قابل ملاحظه و بر طبق مطالعات تجربی ناچیز است. در بقیه متن‌ها بر طبق هر دو ملاک وابستگی مکانی سؤال ناچیز است. در نتیجه این دو متن بر طبق نتایج مطالعات شبیه‌سازی موجب انحراف از تک‌بعدی بودن شده‌اند و وجود وابستگی مکانی بین سؤالات را نشان می‌دهد. نتایج واریانس اثر جنسیت نشان داد جنسیت در متن اول ۹۶ و در متن دوم ۸۳ درصد واریانس وجود وابستگی مکانی بین سؤالات متن‌ها را تبیین می‌کند که نشان دهنده تأثیر زیاد جنسیت بر صفات نهفته سنجش شده توسط این دو متن است.

جدول ۳. شاخص قدرت تمیز سؤالات در مدل‌های پاسخ مدرج شده و دو عاملی نظریه در نمونه کلی و پسران و دختران در متن اول

سؤال	ضرب تمیز			ضرب تمیز متن		
	مدل دو پارامتری			مدل دو عاملی (صفت کلی)		
	کلی	پسر	دختر	کلی	پسر	دختر
۱	۰/۸۵۷ (۰/۸۴)	۰/۱۳ (۰/۱۱۳)	۰/۱۷۴ (۰/۱۱۹)	۰/۸۵۷ (۰/۸۴)	۰/۱۳ (۰/۱۱۳)	۰/۱۷۴ (۰/۱۱۹)
۲	۰/۹۱ (۰/۷۶)	۰/۱۲۸ (۰/۱۱۳)	۰/۱۲ (۰/۱۱۱)	۰/۹۱ (۰/۷۶)	۰/۱۲۸ (۰/۱۱۳)	۰/۱۲ (۰/۱۱۱)
۳	۰/۱۲۶ (۰/۸۱)	۰/۱۸۶ (۰/۱۳۳)	۰/۱۳۵ (۰/۱۱۳)	۰/۱۲۶ (۰/۸۱)	۰/۱۸۶ (۰/۱۳۳)	۰/۱۳۵ (۰/۱۱۳)
۴	۰/۱۲۷ (۰/۷۹)	۰/۱۶۱ (۰/۱۱۵)	۰/۸۲ (۰/۹۹)	۰/۱۲۷ (۰/۷۹)	۰/۱۶۱ (۰/۱۱۵)	۰/۸۲ (۰/۹۹)
۵	۰/۱۲۷ (۰/۷۵)	۰/۱۵۶ (۰/۱۰۷)	۰/۷۸ (۰/۸۵)	۰/۱۲۷ (۰/۷۵)	۰/۱۵۶ (۰/۱۰۷)	۰/۷۸ (۰/۸۵)
۶	۰/۱۰۹ (۰/۷۸)	۰/۳۳۲ (۰/۱۳۱)	۰/۷۷ (۰/۸۷)	۰/۱۰۹ (۰/۷۸)	۰/۳۳۲ (۰/۱۳۱)	۰/۷۷ (۰/۸۷)
۷	۰/۸۳ (۰/۷۲)	۰/۱۱ (۰/۱۰۶)	۰/۱۰۸ (۰/۱۰۲)	۰/۸۳ (۰/۷۲)	۰/۱۱ (۰/۱۰۶)	۰/۱۰۸ (۰/۱۰۲)
۸	۰/۶۷ (۰/۷۳)	۰/۱۳۷ (۰/۱۱۲)	۰/۱۱۶ (۰/۱۱۱)	۰/۶۷ (۰/۷۳)	۰/۱۳۷ (۰/۱۱۲)	۰/۱۱۶ (۰/۱۱۱)
۹	۰/۱۶۲ (۰/۷۷)	۰/۱۶۲ (۰/۱۱۸)	۰/۱۰۴ (۰/۱۰۲)	۰/۱۶۲ (۰/۷۷)	۰/۱۶۲ (۰/۱۱۸)	۰/۱۰۴ (۰/۱۰۲)
۱۰	۰/۵۲ (۰/۷۲)	۰/۹۴ (۰/۱۰۸)	۰/۷۸ (۰/۱۰۶)	۰/۵۲ (۰/۷۲)	۰/۹۴ (۰/۱۰۸)	۰/۷۸ (۰/۱۰۶)

۰/۰۹۳ (۰/۱۳۳)	۰/۱۰۵ (۰/۱۳۹)	۰/۳۷۵ (۰/۱۲۵)	۰/۱۰۷ (۰/۱۰۷)	۰/۱۰۶ (۰/۱۰۶)	۲(۰/۰۹) (۰/۰۹)	۷۲/۰ (۰/۰۹۷)	۷۶/۰ (۰/۱۰۲)	۰/۰۵ (۰/۰۶۹)	۱۱
۰/۱۸۳ (۰/۱۵۲)	۰/۲۰۴ (۰/۱۵۱)	۰/۳۴۷ (۰/۱۱۳)	۳(۰/۱۱) (۰/۱۱۳)	۰/۱۱۷ (۰/۱۱۷)	۰/۰۸۸ (۰/۰۸۸)	۰/۰۵۹ (۰/۰۹۷)	۰/۰۸۷ (۰/۱۰۵)	۰/۰۶۹ (۰/۰۷)	۱۲
۰/۱۲۱ (۰/۱۳۸)	۰/۷۸۸ (۰/۱۴)	۰/۲۲۰ (۰/۰۹۵)	۰/۱۰۹ (۰/۱۰۹)	۰/۱۱۲ (۰/۱۱۲)	۱(۰/۰۸) (۰/۰۸)	۰/۰۵۷ (۰/۰۹۷)	۰/۰۹۸ (۰/۱۰۵)	۰/۰۶۹ (۰/۰۷۲)	۱۳

پارامتر شیب و ضریب تبیین سؤالات در مدل‌های دو پارامتری و دو عاملی (خطای استاندارد اندازه گیری)

جدول ۴. نتایج واریانس‌های تصادفی اثر متن‌ها (آزمونچه) در مطالعه پرلز ایران

واریانس تبیین شده توسط جنسیت	واریانس اثر جنسیت	وابستگی مکانی سؤال		واریانس اثر متن	صفات نهفته
		مطالعات تجربی	مطالعات شبیه سازی		
-	-	ناچیز	ناچیز	۰/۰۹۸**	درک مطلب کلی
۰/۹۶	۰/۰۱۷	ناچیز	زیاد	۰/۴۰۷**	درک مطلب ادبی ۱
۰/۸۳	۰/۰۶۱	ناچیز	زیاد	۰/۳۵۷**	درک مطلب ادبی ۲
-	-	ناچیز	ناچیز	۰/۰۵۴**	درک مطلب ادبی ۳
-	-	ناچیز	ناچیز	۰/۱۲۹**	درک مطلب ادبی ۴
-	-	ناچیز	ناچیز	۰/۰۴**	درک مطلب ادبی ۵
-	-	ناچیز	ناچیز	۰/۰۲	درک مطلب اطلاعاتی ۱
-	-	ناچیز	ناچیز	۰/۰۳۹**	درک مطلب اطلاعاتی ۲
-	-	ناچیز	ناچیز	۰/۲۱۶**	درک مطلب اطلاعاتی ۳
-	-	ناچیز	ناچیز	۰/۲۳۷**	درک مطلب اطلاعاتی ۴

ت: آزمون تفاوت بین میانگین‌های نهفته و واریانس‌های عاملی گروه پسران و دختران. جدول ۵ نتایج مقایسه میانگین و واریانس‌های عاملی صفت کلی و صفات خاص (متن‌ها) در دو گروه پسران و دختران را نشان می‌دهد. در این مطالعه گروه پسران به علت حجم نمونه بالاتر به عنوان گروه مرجع و گروه دختران به عنوان گروه کانونی (مقایسه) قرار گرفت. در این حالت میانگین و واریانس صفات نهفته در گروه مرجع (گروه پسران) جهت معین کردن مدل به ترتیب برابر

صفر و یک قرار داده شد. نتایج نشان داد میانگین صفت کلی در گروه دختران ۰/۰۳۹ انحراف استاندارد پایین‌تر از میانگین صفت کلی پسران است اما به لحاظ آماری معنادار نیست. در سه متن مرتبط با درک مطلب ادبی و سه متن مرتبط با درک مطلب اطلاعاتی دختران عملکرد بالاتری نسبت به پسران دارند و به لحاظ آماری معنادار هستند در بقیه متن‌ها تفاوت معناداری مشاهده نشد. در درک مطلب کلی، دوم متن ادبی و چهار متن اطلاعاتی واریانس عاملی گروه دختران کوچکتر از پسران است در نتیجه سطح پیشرفت در درک مطلب کلی، دو متن ادبی و چهارم متن اطلاعاتی در گروه دختران همگون‌تر از گروه پسران است. در بقیه متن‌ها بین دو گروه تفاوتی معناداری ملاحظه نشد.

جدول ۵. تفاوت بین میانگین‌ها و واریانس‌های عاملی صفات نهفته کلی و خاص در بین پسران و دختران

گروه مقایسه (دختر)					درک مطلب
میانگین (خطا)	واریانس (خطا)	درک مطلب	میانگین (خطا)	واریانس (خطا)	
۰/۰۳۹(۰/۰۵۸)	۰/۵۲۶(۰/۰۶)**	اطلاعاتی ۱	۰/۱۹۱(۰/۰۶۷)**	۰/۸۴(۰/۱۱)	کلی
۰/۲۹۹(۰/۰۶۴)**	۰/۸۲۵(۰/۱۱)	اطلاعاتی ۲	۰/۰۰۷(۰/۰۶۴)	۰/۷۲۲(۰/۱)**	ادبی ۱
۰/۲۴۱(۰/۰۶۸)**	۱/۰۷۸(۰/۱۵)	اطلاعاتی ۳	۰/۰۷۷(۰/۰۶۲)	۰/۷۵۳(۰/۰۹)**	ادبی ۲
۰/۰۵۳(۰/۰۶۵)	۰/۶۷۱(۰/۱)**	اطلاعاتی ۴	۰/۱۳۸(۰/۰۶۸)**	۰/۶۷۳(۰/۰۸)**	ادبی ۳
۰/۲(۰/۰۶۵)**	۰/۸۱۳(۰/۱)	اطلاعاتی ۵	۰/۱۹۳(۰/۰۶۳)**	۰/۶۶۸(۰/۰۸)**	ادبی ۴
۰/۱(۰/۰۶۵)	۰/۶۳۱(۰/۰۹)**	-	-	-	ادبی ۵

د: کارکرد افتراقی سؤال مبتنی بر مدل دو عاملی نظریه سؤال-پاسخ. کارکرد افتراقی سؤالات متن‌های خواندن پرلز ۲۰۱۱ ایران در بین دانش‌آموزان پسر و دختر ایرانی بر حسب رویکرد کای و همکاران (۲۰۱۱) در جدول ۶ آورده شده است (به علت تعداد زیاد جداول نتایج، جدول نتایج متن اول آورده شده است). نتایج نشان داد از ۱۴۶ سؤال ۲۸ سؤال (۱۹/۱۸ درصد) دارای کارکرد افتراقی یکنواخت (یکپارچه)، ۴۶ سؤال (۳۱/۵ درصد) دارای کارکرد افتراقی غیر یکنواخت (غیریکپارچه) و ۷۲ سؤال (۴۹/۳۱ درصد) هم دارای هم‌ارزی سؤال هستند. در سؤالات دارای کارکرد افتراقی یکنواخت، ۹ سؤال (۶/۱۷ درصد) چهارگزینه‌ای و ۱۹ سؤال (۱۳/۰۱ درصد)

پاسخ‌ساز بودند همچنين ۱۵ سؤال (۵۳/۵۷ درصد) در گروه پسران (۵ سؤال چندگزينه‌اي و ۱۰ سؤال پاسخ‌ساز) و ۱۳ سؤال (۴۶/۴۳ درصد) در گروه دختران (۵ سؤال چندگزينه‌اي و ۸ سؤال پاسخ‌ساز) دشوارتر نسبت به گروه ديگر در کل پيوستار توانايي بودند. در سؤالات داراي کارکرد افتراقي غير يکنواخت، ۲۰ سؤال (۱۳/۷ درصد) چهارگزينه‌اي و ۲۶ سؤال (۱۷/۸ درصد) پاسخ-ساز بودند و ۲۱ سؤال در گروه پسران (۵ سؤال چندگزينه‌اي و ۱۶ سؤال پاسخ‌ساز) و ۲۵ سؤال در گروه دختران (۱۴ سؤال چندگزينه‌اي و ۱۱ سؤال پاسخ‌ساز) داراي شيب بيشتری بودند. در ۷۲ سؤال داراي هم‌ارزي سؤال نيز ۴۰ سؤال (۲۷/۴ درصد) چهارگزينه‌اي و ۳۲ سؤال (۲۱/۹۱ درصد) پاسخ‌ساز بودند. در کليه سؤالات نيز ميانه‌گين دختران بالاتر از پسران بود و سطح پيشرفت دختران از همگوني نسبي بيشتری نسبت به پسران برخوردار است.

جدول ۶. کارکرد افتراقي سؤالات خوشه‌هاي سؤال آزمون پرلز ۲۰۱۱ بر حسب جنسيت

سؤال	مقدار G^2	درجه آزادي	وضعيت فرض	گروه مرجع (پسران)		گروه کانوني (دختران)			DIF
				شيب	آستانه اول	آستانه دوم	شيب	آستانه اول	
۱	۳/۹	۲	رد	۱/۳	-۰/۶۱	-	۱/۵۸	-۰/۷	غیریکنواخت
۲	۱/۲	۲	قبول	۰/۹	-۰/۳۷	-	۰/۸۱	-۰/۲۶	خیر
۳	۴/۲	۲	رد	۱/۵۱	-۰/۸۸	-	۱/۸۶	-۰/۹۵	غیریکنواخت
۴	۰/۳	۲	قبول	۱/۲۶	-۱/۰۲	-	۱/۲۴	-۱/۱	خیر
۵	۴/۱	۲	رد	۰/۸۲	-۱/۱۵	-	۰/۹۶	-۱/۳	یکنواخت

DIF	گروه کانونی (دختران)				گروه مرجع (پسران)			برابری همه پارامترها			سؤال	
	انحراف استاندارد	میانگین	آستانه دوم	آستانه اول	شیب	آستانه دوم	آستانه اول	شیب	وضعیت فرض	درجه آزادی		مقدار G^2
یکساخت	۰/۹۲	۰/۱۲	-	۰/۰۴	۱	-	-۰/۳۲	۰/۸۸	رد	۲	۶/۶	۶
غیریکساخت	۰/۹۳	۰/۱۲	۱/۳	۰/۸۹	۱/۰۸	۱/۸۴	۱/۱۷	۰/۸۳	رد	۳	۴/۱	۷
خیر	۰/۹۳	۰/۱۳	-	-۰/۲۶	۱/۲۲	-	-۰/۱۳	۱/۷	قبول	۲	۲/۱	۸
یکساخت	۰/۹۳	۰/۱۳	۰/۲۳	-۰/۵۳	۱/۹۵	۰/۴	-۰/۳۱	۱/۶۷	رد	۳	۹/۹	۹
یکساخت	۰/۹۲	۰/۱۲	-	۱/۲۳	۰/۹۲	-	-۰/۷	۱/۰۹	رد	۲	۸/۳	۱۰
خیر	۰/۹۲	۰/۱۲	-	-۰/۳۲	۰/۵۵	-	-۰/۱۴	۰/۶	قبول	۲	۰/۶	۱۱
غیریکساخت	۰/۹۳	۰/۱۳	۲/۰۶	۰/۸۳	۱/۴۷	۲/۴۵	۱/۱۴	۱/۲۵	رد	۳	۴/۶	۱۲
خیر	۰/۹۳	۰/۱۳	-	-۰/۱۶	۰/۸۳	-	-۰/۳۸	۰/۹۲	قبول	۲	۱/۶	۱۳

بحث و نتیجه‌گیری

این مطالعه با هدف بررسی بعدیت و کارکرد افتراقی سؤالات آزمون مبتنی بر متن پرلز ۲۰۱۱ ایران انجام گرفت. مشکل اصلی آزمون‌های مبتنی بر متن تخطی از مفروضه استقلال مکانی سؤال هنگام کاربرد مدل‌های تک‌بعدی نظریه سؤال-پاسخ به واسطه در نظر نگرفتن وابستگی درونی سؤالات درون هر متن است، در این حالت استفاده از مدل‌های تک‌بعدی موجب برآزش خوب با داده‌ها نمی‌شود و این امر می‌تواند بر دقت برآورد پارامترهای سؤال و توانایی تأثیر بگذارد.

جهت بررسی بعدیت و کارکرد افتراقی سؤالات آزمون‌های مبتنی بر متن امروزه مدل‌های مبتنی بر متن همچون مدل دو عاملی نظریه سوال-پاسخ کاربرد زیادی در سنجش‌های آموزشی و روان-شناختی پیدا کرده است.

نتایج این مطالعه در بررسی بعدیت سؤالات متن‌های خواندن نشان داد هم در نمونه کلی ایران و هم در گروه دانش‌آموزان دختر و پسر، زمانی که مدل تک‌بعدی پاسخ مدرج شده بدون توجه به اثر متن‌ها به کار برده می‌شود اکثر سؤالات دارای شیب‌های خیلی کم روی صفت نهفته هستند (زیر ۰/۸). دی‌مارس (۲۰۱۳) در تفسیر شیب‌های سؤالات در مدل دو عاملی بیان می‌کند شیب‌ها در دامنه ۰ تا ۰/۸ شیب‌های خیلی کم، در دامنه ۰/۸ تا ۱/۴۹ شیب‌های متوسط و ۱/۵ و بالاتر شیب‌های قابل قبول محسوب می‌شوند. معمولاً اگر آزمون تک‌بعدی باشد شیب اکثر سؤالات روی صفت نهفته زیربنایی باید بزرگ باشد و در دامنه ۰/۸ تا ۱/۴۹ و یا ۱/۵ به بالا قرار گیرند اما در سؤالات ایران در هر سه گروه نمونه کلی، پسران و دختران شیب‌ها در دامنه ۰ تا ۰/۸ قرار دارند که شیب‌های خیلی کم محسوب می‌شوند. نتایج بررسی ضریب تبیین سؤالات در مدل پاسخ مدرج شده نیز نشان داد که نسبت واریانس تبیین شده سؤالات در مدل پاسخ مدرج شده توسط صفت نهفته زیربنایی کم است. شیب کم و ضریب تبیین کوچک سؤالات می‌تواند نشان دهنده این باشد علاوه بر صفت نهفته زیربنایی درک مطلب که بر پاسخ به سؤالات در متن‌های خواندن پرلز اثر می‌گذارد عوامل ثانوی دیگر نیز می‌تواند اثرگذار باشند که به عنوان عوامل مزاحم یا صفات ثانویه معرفی می‌شوند. در مطالعه ریجمن (۲۰۱۱) که بر روی متن‌های پرلز ۲۰۰۶ انجام شد نتایج اثراتی از بعدیت در سؤالات را نشان داد.

جهت بررسی تأثیر صفات نهفته ثانویه از مدل دو عاملی استفاده شد. نتایج بررسی شیب‌ها در مدل دو عاملی نشان داد شیب اکثر سؤالات روی صفت نهفته خاصی که سؤال می‌سنجد بزرگتر از صفت نهفته کلی است که با نتایج شیب‌های خیلی کم روی صفت نهفته کلی در مدل پاسخ مدرج شده همسو است، زیرا بارهای عاملی کوچک روی یک صفت نهفته می‌تواند نشان

دهنده وجود عوامل تاثیرگذار دیگر باشد. وجود شیب‌های بزرگ سؤالات روی متن‌ها همچنین این مطلب را که سؤالات با صفت زیربنایی خود نسبت به صفت کلی که توسط تمام سؤالات اندازه‌گیری می‌شود رابطه قوی‌تری دارد تأیید می‌کند. این نتایج نشان داد با استفاده از مدل‌هایی که اثر صفات خاصی را که توسط متن‌ها سنجیده می‌شوند در نظر می‌گیرند می‌توان وابستگی‌های اضافی بین سؤالات را به واسطه همبستگی درونی سؤالات درون هر متن که یک مطلب یا محتوای خاص را می‌سنجند بهتر تبیین کرد. نتایج ضریب تبیین سؤالات نیز نشان داد نسبت واریانس تبیین شده سؤالات در مدل دو عاملی افزایش پیدا کرده است. دی مارس (۲۰۱۳) بیان می‌کند علت این امر این است که در مدل دو عاملی یک عامل کلی وجود دارد که توسط تمام سؤالات اندازه‌گیری می‌شود و علاوه بر عامل کلی چندین عامل خاص وجود دارند که برای تبیین واریانس باقی مانده زیر مجموعه‌ای از سؤالات که در آن اشتراک دارند به کار برده می‌شوند. در این مدل پاسخ‌های سؤال تابعی از عامل کلی و یک عامل خاص هستند زیرا عامل‌های خاص متعامد با عامل کلی هستند و بازنمایی‌کننده واریانس مشترک برای یک گروه از سؤالات فراتر از عامل اندازه‌گیری شده به وسیله مقیاس به عنوان یک کل هستند. همچنین دی مارس (۲۰۰۶) بیان می‌کند از آنجایی که پاسخ به سؤالات درون یک متن با یک صفت ثانویه ارتباط دارند این صفت ثانویه ممکن است دانش پیشین یا مهارت خاص برای متن یا ممکن یک سطح علاقه یا دیگر عوامل انگیزشی خاص برای متن باشد.

جهت تعیین عوامل ثانوی واریانس‌های تصادفی اثر متن‌ها مورد بررسی قرار گرفت. نتایج نشان داد دو متن مرتبط با درک مطلب ادبی (گل‌های روی پشت بام و پرواز کن عقاب، پرواز کن) بر طبق مطالعات شبیه‌سازی دارای اثر متن یا وابستگی مکانی سؤال هستند و موجب انحراف از تک بعدی بودن سؤالات شده‌اند. نتایج نشان داد جنسیت حدود زیادی از واریانس این وابستگی را تبیین می‌کند، در نتیجه این دو متن تا حد زیادی وابسته به جنسیت هستند و عوامل متفاوتی بر عملکرد پسران و دختران در این دو متن تاثیرگذار هستند. ین (۱۹۹۳) و تائو (۲۰۰۸)

به نقل از لرد (۱۹۸۰)، هامبلتون، سوامیناتان و راجرز (۱۹۹۱) بیان می‌کنند اصل اساسی ایجاد وابستگی مکانی سؤال وجود یک بعد اضافی است که به طور پیوسته بر عملکرد بعضی از دانش-آموزان در بعضی سؤالات نسبت به گستره بزرگتر دیگری از سؤالات اثر می‌گذارد. واینر و تیسن (۱۹۹۶)، به نقل از تائو، (۲۰۰۸) این بعد اضافی را بعد تصادفی^۱ می‌نامند که آزمون قصد اندازه-گیری آن را ندارد و در برابر بعد ثابت قرار می‌گیرد که مدل‌های چندبعدي نظريه سوال-پاسخ قصد اندازه‌گیری و برآورد آن را دارند. ین (۱۹۹۳) بیان می‌کند اگر اثر این عامل اضافی ناخواسته ثابت باشد به این معنا است که بر همه دانش‌آموزان و همه سؤالات به روش یکسان اثر می‌گذارد در این حالت وابستگی مکانی سؤال رخ نمی‌دهد. زمانی که اثر ثابت روی می‌دهد هنوز مدل تک‌بعدي نظريه سوال-پاسخ قابل کاربرد است با این استثناء که توانایی زیربنایی برآورده شده برای آزمودنی مجموع ابعاد اولیه و ثانویه خواهد بود اما زمانی که اثر عامل اضافی به طور مختلفی بر دانش‌آموزان معینی یا سؤالات معینی اثر می‌گذارد وابستگی مکانی سؤال در بین سؤالات روی می‌دهد. تعداد زیادی از شرایط زمینه‌ای می‌تواند منجر به رخ داد وابستگی مکانی سؤال شود. ین (۱۹۹۲)، به نقل از تائو، (۲۰۰۸) بیان می‌کند این ابعاد اضافی می‌تواند ناشی از عوامل مختلفی همچون یک عامل خارجی، سرعت، خستگی، شکل سؤال یا پاسخ، زنجیره‌ای بودن سؤالات، رهنمودهای نمره‌گذاری سؤالات و وابستگی متن باشد. ین (۱۹۹۳) و تائو (۲۰۰۸) به نقل از نتایج مطالعات لی^۲ (۲۰۰۴)؛ سیرسی، تیسن و واینر^۳ (۱۹۹۱)؛ تیسن، استینبرگ و مونی^۴ (۱۹۸۹) بیان می‌کنند مهمترین علت وابستگی سؤالات در متن‌های خواندن آزمون‌های درک مطلب، وابستگی متن است که بیشترین تأثیر را بر وابستگی مکانی سؤال در مطالعات بین‌المللی دارد. وابستگی متن زمانی روی می‌دهد که بعضی دانش‌آموزان اطلاعات بیشتر یا علاقه بیشتری به موضوعات معین بیان شده در متن داشته باشند در نتیجه احتمال بیشتری وجود دارد به سؤالات

1. random dimension
2. Lee
3. Sireci, Thissen & Wainer
4. Thissen, Steinberg & Mooney

پاسخ درست دهند. همچنین علت وابستگی می‌تواند وابستگی شکل سؤال یا پاسخ باشد. این وابستگی به علت وجود پاسخ‌های متنوع به سؤالات چندگزینه‌ای و پاسخ‌ساز بر حسب طول یا نوع متفاوت پاسخ است که نشان می‌دهد بعضی از دانش‌آموزان در نوشتن پاسخ، داستان و یا ترسیم یک تصویر بهتر از دانش‌آموزان دیگر هستند. نتایج واریانس تصادفی اثر متن در گروه پسران و دختران نشان داد که واریانس تصادفی دو متن در گروه پسران به ترتیب $0/092$ و $0/118$ و در گروه دختران به ترتیب $0/39$ و $0/47$ است در نتیجه احتمال وابستگی متن یا وجود پراکندگی بیشتر در اطلاعات و علایق دختران و یا دادن پاسخ‌های متنوع (وابستگی سؤال و پاسخ) به سؤالات در بین گروه دختران بیشتر از پسران است؛ بنابراین جنسیت دانش‌آموزان نقش مهمی در تنوع پاسخ‌ها و اطلاعات بیشتر و علایق به خواندن در این دو متن دارد.

در بررسی کارکرد افتراقی سؤالات بین دو گروه دختران و پسران ابتداء پارامتر اثر مورد بررسی قرار گرفت. پارامتر اثر بیان‌کننده عملکرد افتراقی سؤال در بین دو گروه است که از تفاوت در متوسط توانایی‌های دو گروه به وجود می‌آید و منعکس‌کننده این است چگونه دو گروه در سطوح توانایی‌هایشان بر حسب عملکرد در سؤال با یکدیگر متفاوت هستند. نتایج اثر $(0/024)$ نشان داد بین متوسط برآوردهای توانایی پسران و دختران بر حسب سؤالات متن‌های پرلز ایران تفاوت معناداری وجود ندارد و در صورت وجود تفاوت در عملکرد پسران و دختران در سؤالات نمی‌توان آن را به تفاوت در متوسط توانایی دو گروه نسبت داد. در بررسی کارکرد افتراقی سؤالات از شیوه تحلیل کارکرد افتراقی سؤال در مدل دو عاملی استفاده شد. در این مدل به جای نمره کل که در مدل‌های سنتی به عنوان ملاک هم‌تا بودن توانایی نهفته گروه‌ها استفاده می‌شود از شیوه لنگر قرار دادن سؤالات استفاده می‌شود. در مدل دو عاملی طبق رویکرد کای و همکاران (۲۰۱۱) برای بررسی کارکرد افتراقی سؤالات مدل‌های دو عاملی چندبعدی در هر مرحله دو دسته سؤال وجود دارد، سؤالات لنگر و سؤالات داوطلب بررسی کارکرد افتراقی سؤال. سؤالات داوطلب سؤالات متنی هستند که باید مورد بررسی قرار گیرد و سؤالات لنگر

سؤالات متن‌های دیگر هستند. در هر مرحله سؤالات متن مورد بررسی با لنگر قرار دادن سؤالات متن‌های دیگر مورد بررسی قرار می‌گیرد. نتایج تحلیل کارکرد افتراقی سؤالات نشان داد ۲۸ سؤال (۱۹/۱۸ درصد) دارای کارکرد افتراقی یکنواخت، ۴۶ سؤال (۳۱/۵ درصد) دارای کارکرد افتراقی غیریکنواخت و ۷۲ سؤال (۴۹/۳۱ درصد) فاقد کارکرد افتراقی سؤال هستند. زمانی که سؤالات دارای کارکرد افتراقی یکنواخت هستند منحنی ویژگی سؤال در بین دو گروه متفاوت اما موازی است و نشان می‌دهد سؤال در یک گروه نسبت به گروه دیگر در تمام پیوستار توانایی دشوارتر است؛ اما زمانی که سؤال کارکرد افتراقی غیر یکنواخت نشان می‌دهد سؤال در بین دو گروه دارای منحنی ویژگی متفاوت و ناموازی است در نتیجه دو منحنی در یک نقطه با یکدیگر تعامل دارند (وایبرگ^۱، ۲۰۰۷)، یعنی سؤال در یک گروه در سطوح معینی از توانایی و برای گروه دیگر در سطوح معین دیگری از توانایی دشوارتر است. نتایج بررسی نشان داد در سؤالات دارای کارکرد افتراقی یکنواخت بیشتر تفاوت در سؤالات پاسخ‌ساز و در سؤالات دارای کارکرد افتراقی غیریکنواخت تفاوت در سؤالات پاسخ‌ساز کمی نسبت به سؤالات چندگزینه‌ای بیشتر است. در سؤالات دارای کارکرد افتراقی یکنواخت از بین ۲۸ سؤال، ۵ سؤال چندگزینه‌ای در گروه پسران و ۵ سؤال چندگزینه‌ای در گروه دختران و ۱۰ سؤال پاسخ‌ساز در گروه پسران و ۸ سؤال پاسخ‌ساز در گروه دختران دارای آستانه دشواری کمتری هستند که نشان دهنده سخت بودن سؤال برای آن گروه نسبت به گروه دیگر است. این نتایج نشان می‌دهد که بیشتر تفاوت بین پسران و دختران در سؤالات پاسخ‌ساز است تا سؤالات چندگزینه‌ای و سؤالات پاسخ‌ساز تا حدودی برای دختران نسبت به پسران آسان‌تر هستند. در سؤالات دارای کارکرد افتراقی غیریکنواخت ۲۱ سؤال در گروه پسران و ۲۵ سؤال در گروه دختران دارای شیب بیشتری هستند. در گروه پسران از بین ۲۱ سؤال، ۵ سؤال چندگزینه‌ای و ۱۶ سؤال پاسخ‌ساز و در گروه دختران ۱۴ سؤال چندگزینه‌ای و ۱۱ سؤال پاسخ‌ساز دارای شیب بیشتری هستند. همچنین نتایج نشان داد

1. Wiberg

در سؤالات پاسخ‌ساز هر چه طبقات پاسخ کامل‌تر می‌شود آستانه طبقات در گروه دختران افزایش پیدا می‌کند که نشان دهنده این است سؤالات دارای پاسخ کاملتر در گروه دختران بیشتر از پسران هستند. این نتایج می‌تواند نشان دهد که در سؤالاتی که بیشتر با فرایند ذهنی بازشناسی سروکار دارند پسران و در سؤالاتی که با فرایند ذهنی بازسازی و درک عمیق مطلب سروکار دارند دختران عملکرد بهتری دارند. این نتایج می‌تواند تا حدودی گویایی تعامل بین جنسیت و شکل سؤال نیز باشد زیرا دختران در سؤالات پاسخ‌ساز نمره‌گذاری شده به صورت چند ارزشی نسبت به سؤالات پاسخ‌ساز نمره‌گذاری شده به صورت دو ارزشی از پسران نمرات کامل‌تری به دست آورده‌اند. هیده و لین^۱ (۱۹۸۸)، به نقل از لین و و^۲ (۲۰۰۳) با یک فراتحلیل جامع تفاوت‌های جنسیتی در توانایی کلامی را مورد بررسی قرار دادند، از میان ۵۶ مطالعه دایره لغت، شش گزارش تفاوت‌های جنسیتی را به نفع پسران و هشت مطالعه به نفع دختران گزارش کرده‌اند. در کل فراتحلیل تفاوت جنسیتی معناداری در دایره لغات بین پسران و دختران نشان داده نشد. اگر چه همگونی معناداری در اندازه اثر وجود داشت اما از حیث درک مطلب ۵ مطالعه از ۲۱ مطالعه تفاوت معناداری را به نفع مردان در حالی که ۱۰ مطالعه تفاوت معناداری را به نفع زنان نشان دادند. در کل آنها پی‌بردند که زنان برتری نسبی در خواندن، صحبت کردن، نوشتن و توانایی کلامی دارند اما تفاوت آنقدر کم بود که هیده و لین بیان کردند تفاوت جنسیتی در توانایی کلامی وجود ندارد. همچنین کول^۳ (۱۹۹۷)، به نقل از لین و و^۲ (۲۰۰۳) بیان می‌کند در مطالعه خدمات آزمودن آزموشی^۴ (ETS) برتری زبانی برای زنان در طول ۳۰ سال گذشته بدون تغییر مانده است و زنان در توانایی کلامی دامنه‌ای از تفاوت‌های قابل ملاحظه در نوشتن و استفاده از زبان تا تفاوت‌های کوچک در خواندن و استدلال واژگان دارند. نتایج این مطالعه هم نشان داد در سؤالاتی که نیاز به پاسخ‌های عمقی و فهم بهتر مطلب دارند دختران تا حدودی نسبت به پسران

1. Hyde & Linn
2. Lin & Wu
3. Cole
4. educational testing service (ETS)

عملکرد بهتری دارند. همچنین نتایج این مطالعه نشان داد کارکرد افتراقی سؤال در بین متون ادبی و اطلاعاتی بکار برده شده در مطالعه پرلز تقریباً در بین دو نوع متن یکسان است به طوری که از ۷۴ سؤال دارای کارکرد افتراقی سؤال ۳۶ سؤال ادبی و ۳۸ سؤال اطلاعاتی هستند و اکثر سؤالات ادبی و اطلاعاتی برای دختران نسبت به پسران آسانتر هستند. همچنین نتایج این مطالعه نشان داد دختران نسبت به پسران موفقیت نسبی بیشتری در هر دو حوزه ادبی و اطلاعاتی دارند. در مطالعه پرلز متون خواندن برای تجربه سواد (متن ادبی) اغلب از طریق خواندن داستان در حالی که متون خواندن برای کسب و به کارگیری اطلاعات (متن اطلاعاتی) معمولاً با مقالات حاوی اطلاعات مفید و متنهای آموزشی سروکار دارند. متنهايي که به عنوان اطلاعاتی طبقه‌بندی شده‌اند با سؤالاتی درباره اطلاعات موجود در متن همراه هستند و آنهایی که به عنوان ادبی طبقه‌بندی شده‌اند حاوی سؤالاتی هستند که موضوع، نمودار وقایع، شخصیت‌ها و محیط‌ها را نشان می‌دهند (مولیس، مارتین، کندی، ترانگ و سینسبری^۱، ۲۰۰۹).

با وجود نتایج بدست آمده باید این نکته را متذکر شد که خواندن یک فعالیت روان‌شناختی پیچیده است که تحت تأثیر عوامل مختلفی همچون توانایی، سن، تحول و فرهنگ یادگیرنده است و با سطوح متفاوتی از شناخت هم در صوت‌شناسی^۲ و هم در معنا ارتباط دارد. دانش-آموزانی که بتوانند حافظه و منابع شناختی خود را برای تمرکز بر روی ابعاد پردازش زبان همچون لغات جدید و الگوهای جمله در متنها و موضوعاتی که به آنها ارائه می‌شود بهتر بکار بگیرند موفقیت بیشتری در درک مطلب بدست می‌آورند. به‌طور کلی آزمون‌های استاندارد شده بخصوص در حوزه خواندن از متنهايي تشکیل شده‌اند که یک مجموع سؤال با یک متن خواندن در ارتباط هستند. در چنین موقعیت‌هایی پاسخ‌های سؤال ممکن است از سؤالات دیگر متن مستقل نباشد. پیامد برآورد پارامترهای سؤال با مدل‌های نظریه سوال-پاسخ سنتی سوگیری در برآورد پارامترهای سؤال و فرد به‌خصوص پارامتر تمیز سؤال و توانایی است. همچنین پارامترهای سؤال

1. Mullis, Martin, Kennedy, Trong & Sainsbury
2. phonology

تحت شرایط تخطی از مفروضه استقلال مکانی سؤال ممکن است به درستی برآورد نشوند در نتیجه استفاده از مدل‌های سنتی نظریه سؤال-پاسخ یا مدل‌های تک بعدی نظریه سؤال-پاسخ به برآورد دقیق پارامترهای سؤال و در نتیجه برآورد دقیق سطوح توانایی برآورد شده منجر نمی‌شود. امروزه برای حل این مسائل مدل‌هایی همچون مدل دو عاملی نظریه سؤال-پاسخ مطرح شده است که در تحلیل سؤالات آزمون‌های استاندارد شده مبتنی بر متن همچون پرلز و پیزا جهت تحلیل اثرات وابستگی‌های اضافی اعمال شده به واسطه کاربرد متن‌ها به کار برده می‌شوند. همچنین مدل‌هایی جهت آشکارسازی کارکرد افتراقی سؤال مبتنی بر مدل دو عاملی نظریه سؤال-پاسخ مطرح شده است که وابستگی مکانی بین سؤالات را به واسطه استفاده از متن‌ها در هنگام تحلیل کارکرد افتراقی سؤالات نسبت به مدل‌های سنتی نظریه سؤال-پاسخ در نظر می‌گیرد. در این مدل‌ها در صورت برازش داشتن نه تنها پارامترهای سؤال دقیق‌تر برآورد می‌شوند بلکه سطوح توانایی برآورده شده نیز به واسطه سوگیری در برآورد پارامترها همانند آنچه در مدل‌های سنتی نظریه سؤال-پاسخ رخ می‌دهد دچار سوگیری نخواهد شد.

این مطالعه نیز همانند هر مطالعه‌ای دارای محدودیت‌هایی بود که از جمله آنها می‌توان به مدل‌های به کار برده شده اشاره کرد که نیازمند نمونه‌هایی با حجم بالا است. هنگام تحلیل کارکرد افتراقی سؤال نیاز است طبقات جامعه مورد بررسی قرار گیرد و این می‌تواند کارآمدی این روش‌ها را در برآورد پارامترهای سؤال به علت طبقه‌بندی کردن نمونه کلی به زیر طبقاتی و کاهش تعداد افراد در طبقات کاهش دهد. دوم اینکه روش‌های متعددی برای بررسی بعدیت و کارکرد افتراقی سؤالات وجود دارد که بیشتر با داده‌های شبیه‌سازی شده انجام شده است و برای داده‌های واقعی مطالعات بین‌المللی به صورت گزینش از بین سؤالات و بیشتر روی سؤالات دو ارزشی کار شده است و خیلی کم به سؤالات ترکیبی چندگزینه‌ای و پاسخ‌ساز توجه شده است. همچنین از جمله مطالعاتی که می‌توان در این زمینه برای نتیجه‌گیری دقیق‌تر انجام داد انجام یک

مطالعه طولی روی چهار دوره برگزار شده پرلز از ۲۰۰۱ تا ۲۰۱۶ (داده‌های ۲۰۱۶ هنوز منتشر نشده است) جهت بررسی روند وجود سوگیری سازه و سؤال در این مطالعات است.

منابع

- Alper Kose, I., & Demirtasli, N. C. (2012). Comparison of unidimensional and multidimensional models based on item response theory in terms of both variables of test length and sample size. *Procedia- Social and Behavioral Sciences*, 46:135 – 140.
- Cai L, Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, 16: 221–248.
- DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to Testlet-Based tests. *Journal of Educational Measurement*, 43(2):145–168.
- DeMars, C. E. (2013). A Tutorial on interpreting bifactor model scores, *International Journal of Testing*, 13(4): 354-378.
- Devins, G.M., Beiser, M., Dion, R., Pelletier, L.G. & Edwards, R.G. (1997) Cross-cultural measurements of psychological well-being: The psychometric equivalence of Cantonese, Vietnamese, and Laotian translations of the affect balance scale. *American Journal of Public Health*, 87: 794-799.
- Duan, J. C., Hardle, W. K. & Gentle, J. E. (2012). *Handbook of computational finance*. Springer, Berlin-Heidelberg.
- Fukuhara, H. (2009). *A differential item functioning model for testlet-based items using bi-factor multidimensional item response theory model: a bayesian approach*. A dissertation for the degree of Doctor of Philosophy, Department of Educational Psychology and Learning Systems, florida state university libraries
- Fukuhara, H., & Kamata, A. (2011). A bifactor multidimensional item response theory model for differential item functioning analysis on testlet-based items, *Applied Psychological Measurement*, 35(8): 604–622.
- Goldstein, H. (2008). How may we use international comparative studies to inform education policy?, Retrieved Februray 19, 2015, from <http://www.bristol.ac.uk/medialibrary/sites/cmm/migrated/documents/how-useful-are-international-comparative-studies-in-education.pdf>
- Johnson, T. P. (1998). Approaches to equivalence in cross-cultural and cross-national survey research. in J harknes(eds). *Cross-Cultural survey*

- Equivalence* (pp 1-40). ZUMA-nachrichten spezial band 3, Mannheim, ZUMA.
- Kirsch, I., De Jong, J., Lafontaine, D., McQueen, J., Mendelovits, J., & Monseur, C. (2002). *Reading for change: Performance and engagement across countries result from PISA 2000*. New York, Organisation for Economic Cooperation and Development (OECD).
- Lenkeit, J., Chan, J., Hopfenbeck, T. N., & Baird, J. A. (2015). A review of the representation of PIRLS related research in scientific journals. *Educational Research Review*, 16, 102-115. DOI: <http://dx.doi.org/10.1016/j.edurev.2015.10.002>
- Lin, J., & Wu, F. (2003). Differential performance by gender in foreign language testing, The university of Alberta. The centre for research in applied measurement and evaluation. Retrieved February 19, 2015, from <https://pdfs.semanticscholar.org/f0c1/e30566e69e73841a52476cab5c1f5014518f.pdf>
- Ling Ping, H. & Islam, M. A. (2008). Analyzing incomplete categorical data: Revisiting maximum likelihood estimation (MLE) procedure. *Journal of Modern Applied Statistical Methods*, 7 (2): 488-500.
- MD Desa, Z. N. D. (2012). *Bi-factor multidimensional item response theory modeling for subscores estimation, reliability, and classification*. A dissertation for the degree of PHD in Department of Psychology and Research in Education, the University of Kansas.
- Montgomery, M. (2014). *Unidimensional models do not fit unidimensional mixed format data better than multidimensional models*. A dissertation for the degree of PHD in Department of Psychology and Research in Education, University of Kansas.
- Mullis, I. V. S., Martin, M.O., Kennedy, A. M., Trong, K. L., & Sainsbury, M. (2009). *PIRLS 2011 assessment framework, the international association for the evaluation of educational achievement (IEA)*. Amsterdam, the Netherlands, TIMSS & PIRLS international study center lynch school of education, Boston College.
- Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling*, 52 (4): 354-379.
- Ravand, H. (2015). Assessing testlet effect, impact differential testlet, and item functioning using cross-classified multilevel measurement modeling. Retrieved June 20, 2017, from <http://journals.sagepub.com/doi/pdf/10.1177/2158244015585607>

- Rijmen, F. (2011). Hierarchical factor item response theory models for PIRLS: capturing clustering effects at multiple levels. *IERI monograph series: issues and methodologies in large-scale assessments*, 4: 59-74
- Suksuwan, s., Junpeng, P., Ngudgratoke, S., & Guayjarernpanishk., P. (2012). The effect of proportion common items with mixed format test on multidimensional item response theory linking. *procedia-social and behavioral sciences*, 69: 1505-1511.
- Tao, W. (2008). *Using the score-based testlet method to handle local item dependence*. A dissertation for the degree of PHD in department of Educational Research, Measurement, and Evaluation, Boston College
- Van de Vijver, F. J. R. & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage Publications
- 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.
- Yen, W. M. (1993). Scaling performance assessment: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187-213.
- Zenisky, A. P., Hambleton, R. K., & Sireci, S. G. (2001). Effects of local item dependence on the validity of IRT item, test, and ability statistics. *MCAT Monograph*. Association of American Medical Colleges, Section for the Medical College Admission Test Monograph N. 5.

Archive SID