

## تحلیل بعدیت، اثر متن، تأثیر و کارکرد افتراقی سؤال در آزمون‌های مبتنی بر متن

محمد احمدی ده قطب‌الدینی\*

ابراهیم خدایی\*\*

ولی‌الله فرزاد\*\*\*

علی مقدم‌زاده\*\*\*\*

مسعود کبیری\*\*\*\*\*

### چکیده

این مطالعه با هدف بررسی اثر تک‌زبانه، دوزبانه بودن دانش‌آموزان ایرانی بر بعدیت، وابستگی مکانی سؤال، تأثیر و کارکرد افتراقی سؤال متن‌های مطالعه پرلز ۲۰۱۱ انجام گرفت. بعدیت با مقایسه مدل تک‌بعدی پاسخ مدرج شده با مدل دو‌عاملی چندبعدی نظریه سؤال-پاسخ، تأثیر و وابستگی مکانی سؤال با استفاده از مدل دوسطحی دو‌عاملی و کارکرد افتراقی سؤال با مدل دو‌عاملی چندگروهی کای و همکاران (۲۰۱۱) بررسی شد. نتایج بعدیت نشان داد مدل دو‌عاملی نسبت به مدل پاسخ مدرج شده برازندگی بهتری با داده‌ها دارد و وابستگی مکانی سؤال بین سؤال‌های دو صفت ادبی موجب انحراف از تک‌بعدی بودن شده است، تفاوت زبانی نیز مقدار زیادی از واریانس آنها را تبیین می‌کند. نتایج تأثیر نشان داد متوسط برآورد توانایی‌های تک‌زبان‌ها بالاتر از دوزبان‌ها است. همچنین نتایج نشان داد سؤال‌ها با کارکرد افتراقی یکنواخت برای تک‌زبان‌ها دشوارتر از دوزبان‌ها هستند و در سؤال‌های چندگزینه‌ای با کارکرد افتراقی غیر یکنواخت تک‌زبان‌ها عملکرد بهتری دارند. به‌طور کلی، نتایج نشان داد صفات مرتبط با دو متن در بین دانش‌آموزان تک‌زبان و دوزبان به گونه متفاوتی ادراک می‌شوند و در بین دوزبان‌ها وابستگی مکانی سؤال بیشتر از تک‌زبان‌ها است. همچنین نتایج حاکی از تفاوت عملکرد دانش‌آموزان تک‌زبان - دوزبان در سؤال‌ها با شکل آمیخته است.

**واژگان کلیدی:** مدل پاسخ مدرج شده، مدل دو‌عاملی، بعدیت، کارکرد افتراقی سؤال، آزمون‌های مبتنی بر متن

\*دانشجوی دکتری سنجش و اندازه‌گیری، گرایش سنجش آموزش، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه تهران، ایران

\*\*دانشیار گروه روش‌ها و برنامه‌های آموزشی و درسی، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه تهران، ایران (نویسنده مسئول: Khodaie@ut.ac.ir)

\*\*\*دانشیار گروه روان‌شناسی، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه خوارزمی، ایران  
\*\*\*\*استادیار گروه روش‌ها و برنامه‌های آموزشی و درسی، دانشکده روان‌شناسی و علوم تربیتی، دانشگاه تهران، ایران

\*\*\*\*\*عضو هیئت علمی پژوهشگاه مطالعات آموزش و پرورش وابسته به سازمان پژوهش و برنامه‌ریزی آموزشی، ایران

## مقدمه

در سال‌های اخیر، پیمایش‌های بین‌المللی به منظور مقایسه پیشرفت یادگیری دانش‌آموزان در حوزه‌های ریاضی، علوم و خواندن کشورهای مختلف جهان ظهور کرده‌اند. از جمله این پیمایش‌ها می‌توان به مطالعه بین‌المللی روندها در آموزش ریاضی و علوم<sup>۱</sup> (تیمز)، برنامه سنجش بین‌المللی دانش‌آموزان<sup>۲</sup> (پیزا) و مطالعه بین‌المللی پیشرفت سواد خواندن<sup>۳</sup> (پرلز) اشاره کرد. مطالعه پرلز و دیگر مطالعات بین‌المللی، آزمون‌های روان‌شناختی و آموزشی هستند که برحسب محتوا، شکل سؤال، نمره‌گذاری، هدف‌ها و چهارچوب‌های به کار برده شده با یکدیگر متفاوت هستند اما مورد مشترک در میان همه این آزمون‌ها زبان است. زبان به‌عنوان یک ابزار نباید به‌جز در سنجش شایستگی کلامی، تأثیرات افتراقی روی سنجش داشته باشد و سطح زبان مورد نیاز برای یک آزمون باید در بین گروه‌های مختلف دانش‌آموز یکسان باشد. این اصل اساسی یعنی همگونی زبانی<sup>۴</sup> ممکن است در زمینه‌های مختلف زبانی که زبان خانواده و زبان آموزش، متفاوت است تحت تأثیر قرار گیرد و هم‌ارزی روان‌سنجی بین شرح‌های زبانی را به خطر بیندازد (ایلوسو اولیدن و موجیکا لیزاسو، ۲۰۱۴)، این در حالی است که پیش‌نیاز اصلی مقایسه‌های روا بین کشورها و گروه‌های مختلف درون یک جامعه یکسان، هم‌ارزی<sup>۵</sup> (معادل بودن) داده‌های گردآوری شده از محیط‌های مختلف فرهنگی و اجتناب از سوگیری<sup>۶</sup> است (ون‌د و یجور و لئونگ، ۱۹۹۷، به نقل از هالزموئر، سین و نیجسن، ۲۰۰۲). در استاندارد ۲-۳ از استانداردهای آزمون‌های تربیتی و روان‌شناختی<sup>۱۰</sup> (۲۰۱۴)، به نقل از کنسرسیوم سنجش متوازن هوشمند<sup>۱۱</sup> (۲۰۱۵) نیز بیان شده است که توسعه‌دهندگان آزمون باید برای اندازه‌گیری سازه مورد نظر، تأثیر عوامل بالقوه و غیر مرتبط با سازه مورد نظر همچون

1. Trends in International Maths and Science Study (TIMSS)

2. Programme for International Student Assessment (PISA)

3. Progress in International Reading Literacy Study (PIRLS)

4. linguistic homogeneity

5. Elosua Oliden & Mujika Lizaso

6. equivalence

7. bias

8. Van de Vijver & Leung

9. Holzmueller, Singh & Nijssen

10. standards for educational and psychological testing

11. Smarter Balanced Assessment Consortium

ویژگی‌های زبانی، ارتباطی، شناختی، فرهنگی و فیزیکی یا دیگر ویژگی‌ها را به حداقل برسانند.

نتایج مطالعه پرلز ایران در سه دوره ۲۰۰۱، ۲۰۰۶ و ۲۰۱۱ در ارتباط با مقایسه عملکرد خواندن دانش‌آموزان تک‌زبانه و دوزبانه نشان می‌دهد در پرلز ۲۰۰۱ تک‌زبانها و دوزبانها به ترتیب دارای متوسط نمره ۴۴۳ و ۳۹۴ (مولیس، مارتین، گونزالیز، کینیدی<sup>۱</sup>، ۲۰۰۳)، در پرلز ۲۰۰۶، دارای متوسط نمره ۴۴۷ و ۴۲۳ (مولیس، مارتین، کینیدی و فوی<sup>۲</sup>، ۲۰۰۷) و در پرلز ۲۰۱۱ دارای متوسط نمره ۴۷۰ و ۴۰۸ هستند (مولیس، مارتین، فوی و دراگر<sup>۳</sup>، ۲۰۱۲). این نتایج تقریباً در بیشتر کشورهای شرکت‌کننده در این سه دوره نیز صادق است و تک‌زبانها نسبت به دوزبانها عملکرد بهتری در متن‌های خواندن پرلز دارند؛ این در حالی است که نتایج مطالعات نشان می‌دهد دوزبانها در برخی ابعاد همچون کنترل اجرایی<sup>۴</sup>، آگاهی فرازبانی<sup>۵</sup>، حافظه فعال<sup>۶</sup>، آگاهی فراشناختی<sup>۷</sup> و تفکر واگرا<sup>۸</sup> نسبت به تک‌زبانها عملکرد بهتری دارند (مورتن و هارپر<sup>۹</sup>، ۲۰۰۷، به نقل از سیاهابودین<sup>۱۰</sup>، ۲۰۱۳). مسئله‌ای که پاسخ مشخصی برای آن وجود ندارد این است که عملکرد ضعیف دانش‌آموزان دوزبانه ایرانی نسبت به دانش‌آموزان تک‌زبانها در متن‌های مطالعه پرلز یک علت حقیقی مربوط به ساختار نظام آموزشی است یا متن‌های به کار برده شده علی‌رغم اینکه با همکاری کشورهای شرکت‌کننده ساخته شده است به دلایل مختلفی همچون شکل سؤال‌ها<sup>۱۱</sup> (سؤال چندگزینه‌ای<sup>۱۲</sup> در مقابل سؤال پاسخ‌ساز<sup>۱۳</sup>)، استفاده از متن (آزمونچه<sup>۱۴</sup>)، ادراک متفاوت از هدف‌های خواندن و فرایندهای آن در بین گروه‌های

1. Mullis, Martin, Gonzalez & Kennedy

2. Mullis, Martin, Kennedy & Foy

3. Mullis, Martin, Foy & Drucker

4. executive control

5. metalinguistic awareness

6. working memory

7. metacognitive awareness

8. divergent thinking

9. Morton & Harper

10. Syahabuddin

11. item format

12. multiple choice item

13. constructed response item

14. testlet

مختلف یا تفاوت‌های زبانی بین دانش‌آموزان و ... موجب عملکرد متفاوت دانش‌آموزان تک‌زبانه و دوزبانه ایرانی شده است.

نخستین مسئله که در مورد مطالعه پرلز باید به آن توجه کرد این است که اگرچه سنجش جامعی از مهارت‌ها یا فرایندهای خواندن دانش‌آموزان پایه چهارم ابتدایی همچون تمرکز و بازیابی اطلاعاتی بیان شده به‌طور صریح، استنباط‌های مستقیم، تفسیر و تلفیق ایده‌ها و اطلاعات و بررسی و ارزیابی محتوا، زبان و عناصر متنی به دست می‌دهد اما همان‌طور که ریجمن<sup>۱</sup> (۲۰۱۱) بیان کرده است در مطالعه پرلز پرسش‌های سواد خواندن در سطوح چندگانه‌ای خوشه‌بندی شده‌اند. مطالعه پرلز شامل بلوک‌هایی (آزمونچه‌ها یا متن‌ها) از سؤال است. سؤال‌های هر بلوک (آزمونچه) در درون چهار فرایند اصلی خواندن آشیانه کرده‌اند. هر یک از این فرایندهای خواندن نیز در درون دو هدف اصلی خواندن یعنی خواندن به‌منظور اکتساب و استفاده از اطلاعات و خواندن برای تجربه سواد آشیانه کرده‌اند. در نهایت همه ابعاد سواد خواندن درون عامل اصلی درک مطلب آشیانه کرده‌اند. مور<sup>۲</sup> (۲۰۱۵) به نقل از نتایج مطالعات کیم و کولن<sup>۳</sup> (۲۰۰۶) و یائو<sup>۴</sup> (۲۰۰۸) بیان کرده که احتمال چندبعدی بودن آزمون‌ها با شکل آمیخته به واسطه تفاوت در شکل سؤال‌ها، تفاوت در ابعاد محتوایی درون متن‌ها (آزمونچه‌ها) و تفاوت بین آزمون‌ها به واسطه تغییر خرده آزمون‌ها در بین حیطه‌های محتوایی وجود دارد. آلپرکوز و دمیرتاسلی<sup>۵</sup> (۲۰۱۲) بیان کرده‌اند خطر تحلیل داده‌های چندبعدی به صورت تک‌بعدی این است که می‌تواند به نمره خطای بیشتر در برآورد پارامترهای توانایی و سؤال از یک سو و کاهش دقت اندازه‌گیری و برازش داده‌ها با مدل از سوی دیگر منجر شود که نتیجه نهایی آن سوگیری نتایج به دست آمده خواهند بود.

موضوع دیگر در ارتباط با آزمون‌های مبتنی بر متن<sup>۶</sup> همچون پرلز همان‌طور که فوکوهارا<sup>۷</sup> (۲۰۰۹) بیان کرده این است که در آزمون‌های مبتنی بر متن (آزمونچه) سؤال‌های درون یک متن ممکن است به‌طور کامل مستقل از یکدیگر نباشند زیرا

1. Rijmen

2. Moore

3. Kim & Kolen

4. Yao

5. Alper Kose & Demirtasli

6. testlet based tests

7. Fukuhara

سطح فهم متن خواندن ممکن به وسیله دانش و علاقه آزمودنی‌ها درباره محتوای متن خواندن تحت تأثیر قرار گیرد، این امر موجب بروز وابستگی مکانی<sup>۱</sup> بین سؤال‌های متن و در نتیجه تخطی از مفروضه مهم نظریه‌های اندازه‌گیری یعنی استقلال مکانی سؤال<sup>۲</sup> می‌شود. استفاده از متن موجب شده است تا در مدل‌های جدید تحلیل سؤال همچون مدل دوعاملی نظریه سؤال-پاسخ<sup>۳</sup> یک پارامتر اضافی برای اثر متن<sup>۴</sup> (آزمونچه) به مدل سنتی نظریه سؤال-پاسخ اضافه شود. در این مدل‌ها اثر متن (آزمونچه) به‌عنوان یک متغیر تصادفی در نظر گرفته می‌شود تا بتوان وجود یا نبود وابستگی مکانی بین سؤال‌ها را بررسی کرد. تائو<sup>۵</sup> (۲۰۰۸) به نقل از لرد<sup>۶</sup> (۱۹۸۰)، هامبلتون، سوامیناتان و راجرز<sup>۷</sup> (۱۹۹۱) و یین<sup>۸</sup> (۱۹۹۳) اظهار داشته که اصل اساسی ایجاد وابستگی مکانی سؤال، وجود یک بعد اضافی است که به‌طور پیوسته بر عملکرد بعضی از دانش‌آموزان در بعضی سؤال‌ها نسبت به گستره بزرگ‌تر سؤال‌ها اثر می‌گذارد. واینر و تیسسن<sup>۹</sup> (۱۹۹۳)، به نقل از تائو، (۲۰۰۸) این بعد اضافی را بعد تصادفی<sup>۱۰</sup> نامیده‌اند؛ یک بعد مزاحم که آزمون قصد اندازه‌گیری آن را ندارد و در برابر بعد ثابت قرار می‌گیرد که در مدل‌های چندبعدي نظریه سؤال-پاسخ قصد اندازه‌گیری و برآورد آن وجود دارد. تائو (۲۰۰۸) این نوع وابستگی را وابستگی ترکیب<sup>۱۱</sup> می‌داند و در موقعیتی به وجود می‌آید که سؤال‌ها به یک محرک یکسان وابسته باشند (مثل سؤال‌های مربوط به یک متن خواندن در آزمون درک مطلب).

مسئله دیگر در ارتباط با خواندن همان‌طور که راتوچ و هارتیگ<sup>۱۲</sup> (۲۰۱۰) بیان کرده‌اند شناسایی تأثیر متغیرهای گروه‌بندی شده‌ای همچون جنسیت، زبان خواندن در خانه و ناحیه‌ای که مدرسه در آن قرار دارد بر عملکرد یادگیرندگان در آزمون‌های خواندن است. در گزارش پیزا (۲۰۰۶) بیان شده است آزمون‌دهندگان که به زبانی

1. local item dependence

2. local item independence

3. bifactor item-response theory model

4. testlet effect

5. Tao

6. Lord

7. Hambleton, Swaminathan & Rogers

8. Yen

9. Wainer & Thissen

10. random dimension

11. combination dependence

12. Rauch & Hartig

غیر از زبان آزمون در خانه صحبت می‌کنند انتظار می‌رود عملکرد پایین‌تری در سنجش خواندن داشته باشند نسبت به آزمون‌دهندگان که از زبان آزمون برای برقراری ارتباط در خانه استفاده می‌کنند. روند<sup>۱</sup> (۲۰۱۵) بیان کرده که اثر عامل‌های گروه‌بندی فرد همچون جنسیت یا زبان را می‌توان با تأثیر<sup>۲</sup> و کارکرد افتراقی سؤال<sup>۳</sup> و اثر عامل‌های گروه‌بندی سؤال را می‌توان با بررسی اثر متن (آزمونچه) بررسی کرد. به گفته فوکوهارا (۲۰۰۹) کارکرد افتراقی سؤال، شواهد آماری از عملکرد متفاوت در یک سؤال بین حداقل دو گروه نظیر پسران در مقابل دختران، گروه‌های درون جامعه (اقلیت در مقابل اکثریت) و زبان (مثل تک‌زبان در مقابل دوزبان) بعد از کنترل توانایی را نشان می‌دهد. کارکرد افتراقی سؤال زمانی روی می‌دهد که احتمال پاسخ درست به یک سؤال در بین گروه‌ها در صفت نهفته مورد نظر که سؤال آن را می‌سنجد، متفاوت باشد یا از دیدگاه نظریه سؤال-پاسخ<sup>۴</sup> زمانی که گروه‌های مختلف منحنی‌های ویژگی سؤال<sup>۵</sup> متفاوت داشته باشند. تأثیر نیز به تفاوت در توانایی‌های فرد به‌عنوان تابعی از بعضی توصیف‌کننده‌های سطح فرد مانند جنسیت و زبان اشاره دارد (بریتواس، کاوسون، لوک هارت و کایه<sup>۶</sup>، ۲۰۱۲؛ به نقل از روند، ۲۰۱۵). در واقع تأثیر، عملکرد افتراقی در سؤال در بین دو گروه است که از تفاوت در متوسط توانایی‌های دو گروه به وجود می‌آید و انعکاسی از تفاوت دو گروه در سطح توانایی‌هایشان برحسب عملکرد در سؤال است. تفاوت بین تأثیر و کارکرد افتراقی سؤال این است که تأثیر نشان‌دهنده این است چگونه دو گروه در سطوح توانایی‌هایشان برحسب عملکرد در سؤال متفاوت هستند در حالی که کارکرد افتراقی سؤال نشان می‌دهد چگونه افراد با سطح یکسانی از توانایی به‌طور متفاوتی در سؤال عمل می‌کنند (فوکوهارا، ۲۰۰۹). به عبارت دیگر، تأثیر منعکس‌کننده تفاوت در توانایی‌های فرد به‌عنوان تابعی از توصیف‌کننده سطح فرد همچون جنسیت است در حالی که کارکرد افتراقی سؤال به‌عنوان تابعی از توصیف‌کننده سؤال (همچون جنسیت) بعد از کنترل تأثیر است (بریتواس و والکر<sup>۷</sup>، ۲۰۱۲).

1. Ravand

2. impact

3. Differential Item Functioning (DIF)

4. Item Response Theory (IRT)

5. Item Characteristic Curves (ICCs)

6. Beretvas, Cawthon, Lockhart & Kaye

7. Beretvas & Walker

به منظور بررسی اثر متن‌ها بر بعدیت، وابستگی‌های مکانی سؤال، تأثیر و کارکرد افتراقی سؤال در آزمون‌های مبتنی بر متن، در طی سال‌های اخیر تلاش‌هایی برای تدوین مدل‌های ترکیبی با ترکیب دو مکتب فکری نظریه دو عاملی اسپیرمن<sup>۱</sup> و ساختار ساده<sup>۲</sup> ثرستون<sup>۳</sup> (۱۹۲۵) از عوامل چندگانه توانایی‌های ذهنی نخستین<sup>۴</sup> با نظریه سؤال-پاسخ صورت گرفته که به تدوین مدل‌های مختلف چندبعدی در حوزه نظریه سؤال-پاسخ منجر شده است (ام‌دی دیسا<sup>۵</sup>، ۲۰۱۲). از جمله مدل‌های چندبعدی که در سال‌های اخیر در سنجش‌های آموزشی کاربرد زیادی پیدا کرده است مدل دو عاملی چندبعدی نظریه سؤال-پاسخ<sup>۶</sup> است. در مدل دو عاملی نظریه سؤال-پاسخ که توسط گیبسون و هیدیکر<sup>۷</sup> (۱۹۹۲)، به نقل از کای، یانگ و هانسن<sup>۸</sup>، (۲۰۱۱) مطرح شده است یک بعد کلی<sup>۹</sup> و چند بعد اختصاصی<sup>۱۰</sup> (بسته به عامل‌های مورد بررسی مانند شکل سؤال‌ها، متن‌ها و فرایندهای پردازش) وجود دارد. در این مدل هر سؤال دو شیب، یکی بر روی بعد کلی و یکی بر روی بعد اختصاصی خاص خود دارد. در این مدل برای تحلیل سؤال‌های دوارزشی از تابع پیوند لوجیت<sup>۱۱</sup> و برای سؤال‌های چندارزشی از تابع لوجیت تراکمی<sup>۱۲</sup> استفاده می‌شود (کای و همکاران، ۲۰۱۱).

مطالعاتی که از مدل دو عاملی برای بررسی اثر متن در آزمون‌های خواندن استفاده کرده باشد به جز مطالعه ریجمن (۲۰۱۱) که روی تمام داده‌های کشورهای شرکت‌کننده مطالعه پرلز ۲۰۰۶ انجام گرفته است، مطالعه دیگری در ارتباط با آزمون‌های مبتنی بر متن مشاهده نشد. در مطالعه ریجمن نیز سؤال‌ها به صورت دوارزشی تحلیل شده‌اند و به شکل آمیخته سؤال‌ها توجه نشده است. نتایج این مطالعه نشان داد مدل دو عاملی نسبت به مدل دو پارامتری تک‌بعدی با داده‌ها برازش بهتری دارد و اثراتی از تأثیر متن‌ها در مطالعه پرلز وجود دارد. همچنین بین سؤال‌های متن‌ها

1. Spearman's "two-factor" theory

2. simple structure

3. Thurston

4. theory of primary mental abilities

5. MD Desa

6. bi-factor multidimensional IRT model

7. Gibbson & Hedeker

8. Cai, Yang & Hansen

9. general dimension

10. specific dimension

11. logit link function

12. cumulative logit link function

وابستگی‌های مکانی مثبت و منفی وجود دارد. از سوی دیگر، مطالعاتی که داده‌های پرلز را برحسب مدل دوعاملی از جنبه‌های هم‌ارزی سازه و سؤال و با توجه به عامل‌هایی همچون جنسیت، زبان، وضعیت اجتماعی-اقتصادی و ... بررسی کرده باشند نیز مشاهده نشد. با توجه به تأثیر زبان آزمون به‌خصوص در آزمون‌های خواندن بر عملکرد شرکت‌کنندگان یک آزمون و نبود مطالعاتی که برحسب مدل دوعاملی، اثر زبان را بعدیت، وابستگی مکانی سؤال، تأثیر و کارکرد افتراقی سؤال‌های متن‌های مطالعه پرلز بررسی کرده باشد، این مطالعه با توجه به زبان (تک‌زبان-دوزبان بودن) دانش‌آموزان ایرانی شرکت‌کننده در مطالعه پرلز ۲۰۱۱ سؤال‌های زیر را مدنظر قرار داد:

۱- آیا مدل دوعاملی نظریه سؤال-پاسخ نسبت به مدل تک‌بعدی پاسخ مدرج شده برازندگی بهتری با صفات زیربنایی سؤال‌های متن‌های خواندن مطالعه پرلز ۲۰۱۱ ایران دارد؟ (بعدیت) ۲- کدام صفات نهفته خاص مرتبط با متن‌های خواندن مطالعه پرلز ۲۰۱۱ ایران موجب انحراف از تک‌بعدی بودن شده است؟ (اثر متن) ۳- تفاوت زبانی بین دانش‌آموزان چند درصد از واریانس اثر متن‌هایی را که موجب انحراف از تک‌بعدی بودن شده‌اند، تبیین می‌کند؟ ۴- آیا بین متوسط برآورد پارامتر توانایی دانش‌آموزان تک‌زبان و دوزبان تفاوت وجود دارد؟ (تأثیر) ۵. آیا سؤال‌های متن‌های پرلز ۲۰۱۱ در بین دانش‌آموزان تک‌زبان و دوزبان ایرانی دارای پارامترهای سؤال (ضریب تمیز، آستانه دشواری) متفاوتی هستند؟

### روش

این پژوهش، مطالعه‌ای ثانویه است که روی داده‌های پیمایش بین‌المللی پرلز ۲۰۱۱ ایران به‌عنوان یکی از ۴۸ کشور شرکت‌کننده در مطالعه پرلز ۲۰۱۱ انجام گرفت. نمونه شرکت‌کننده در مطالعه پرلز، دانش‌آموزان پایه چهارم ابتدایی هستند. به‌منظور تعیین تک‌زبان و دوزبان بودن دانش‌آموزان ایرانی با توجه به اینکه نظام آموزش رسمی کشور ایران زبان فارسی است و آزمون پرلز ایران به زبان فارسی اجرا شده است، دانش‌آموزانی که در پاسخ به سؤال «آیا زبان آزمون همان زبانی است که در خانه با آن صحبت می‌کنند»، پاسخ «همیشه» را انتخاب کرده بودند به‌عنوان تک‌زبان و دانش‌آموزانی که «گاهی اوقات» و «هرگز» را انتخاب کرده بودند به‌عنوان دوزبان در نظر گرفته شدند. دوزبان‌ها در نمونه ایرانی اغلب به زبان‌های ترکی، کردی، بلوچی،



عرب و دیگر زبان‌ها تکلم می‌کنند. از بین ۵۷۵۸ شرکت‌کننده ایرانی، ۳۲۳۸ نفر (۵۶/۲۳ درصد) تک‌زبان، ۲۴۶۹ نفر (۴۲/۸۷ درصد) دوزبان و ۵۱ نفر (۹ درصد) نیز وضعیت زبان خود را مشخص نکرده بودند که از تحلیل کنار گذاشته شدند.

### ابزار اندازه‌گیری مطالعه پرلز

آزمون سواد خواندن مطالعه پرلز ۲۰۱۱ دارای ۱۰ متن است که در ۱۳ دفترچه توزیع شده‌اند. 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 مجموع ۵ متن ادبی و ۵ متن اطلاعاتی در مطالعه پرلز وجود دارد. هر متن، سؤال‌هایی به شکل چندگزینه‌ای و پاسخ‌ساز دارد؛ در ۱۰ متن ۱۴۶ سؤال مطرح شده است که ۶۲ سؤال (۴۲/۴۷ درصد) چندگزینه‌ای و ۸۴ سؤال (۵۷/۵۳ درصد) پاسخ‌ساز هستند. همچنین از ۱۴۶ سؤال، ۷۵ سؤال (۵۱/۳۷ درصد) ادبی و ۷۱ سؤال (۴۱/۶۳ درصد) اطلاعاتی هستند. جدول (۱) توزیع متن‌های مطالعه پرلز ۲۰۱۱ را برحسب نام متن، تعداد سؤال‌ها، نوع متن و نوع سؤال نشان می‌دهد.

جدول (۱) توزیع متن‌های مطالعه پرلز برحسب تعداد سؤال، نوع متن و شکل سؤال‌ها

دفترچه	نام متن	تعداد سؤال	نوع متن	نوع سؤال	
				چهارگزینه‌ای	پاسخ‌ساز
۱	گل‌های روی پشت بام	۱۳	ادبی	۷	۶
۲	پرواز کن عقاب، پرواز کن	۱۲	ادبی	۷	۵
۳	کهربای درخشان	۱۴	ادبی	۸	۶
۴	گلدان خالی	۲۰	ادبی	۱۱	۹
۵	لئوناردو داوینچی	۱۲	اطلاعاتی	۶	۶
۶	لذت پیاده‌روی روزانه	۱۲	اطلاعاتی	۷	۵
۷	کوسه‌ها	۱۲	اطلاعاتی	۶	۶
۸	در جستجوی عسل	۱۶	اطلاعاتی	۷	۹
۹	کلوچه دشمن	۱۶	ادبی	۷	۹
۱۰	راز دندان غول	۱۹	اطلاعاتی	۸	۱۱

## روش تحلیل

به‌منظور تعیین بعدیت داده‌ها همان‌طور که دی‌مارس<sup>۱</sup> (۲۰۱۳) مطرح کرده است باید شاخص‌های برازندگی دو مدل تک‌بعدی دو پارامتری لجستیک برای داده‌های دوارزشی و مدل پاسخ مدرج شده یا مدل اعتبار سهمی تعمیم‌یافته برای داده‌های چندارزشی با شاخص‌های مدل دوعاملی نظریه سؤال-پاسخ مقایسه شود. از آنجایی که سؤال‌های متن‌های پرلز به شکل آمیخته هستند، مدل تک‌بعدی پاسخ مدرج شده به‌منظور بررسی تک‌بعدی بودن داده‌ها و مدل دوعاملی به‌منظور تعیین اثر متن‌ها بر انحراف از تک‌بعدی بودن استفاده شد. به‌منظور مقایسه این مدل‌ها از آزمون لگاریتم نسبت درست‌نمایی<sup>۲</sup>، ملاک اطلاعات آکائیک<sup>۳</sup> (AIC) و ملاک اطلاعات بیزی<sup>۴</sup> (BIC) استفاده می‌شود. دوآن، هاردل و جیتل<sup>۵</sup> (۲۰۱۲) بیان کرده‌اند که 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 که جریمه‌های متنوعی برای تعداد پارامترهای برآورده شده در نظر می‌گیرند. مدلی که ملاک اطلاعات آکائیک و ملاک اطلاعات بیزی کوچک‌تری داشته باشد، برازندگی بهتری با داده‌ها دارد.

به‌منظور تعیین اینکه کدام متن‌ها موجب انحراف از تک‌بعدی بودن شده‌اند یا اینکه کدام متن‌ها دارای وابستگی مکانی سؤال هستند و تک‌زبان-دوزبان بودن دانش‌آموزان چند درصد از واریانس این وابستگی را تبیین می‌کند از مدل دوسطحی دوعاملی استفاده شد. در مدل دوسطحی دوعاملی دو نوع واریانس تصادفی درون‌گروهی (واریانس در مدل بدون پیش‌بینی کننده) و بین‌گروهی (واریانس در مدل با پیش‌بینی کننده) برای هر یک از متن‌ها وجود دارد. واریانس درون‌گروهی برای تعیین اثر متن یا وجود وابستگی مکانی بین سؤال‌های متن‌ها و واریانس بین‌گروهی برای کمک به تعیین میزان تأثیر متغیر بین‌گروهی (در این مطالعه زبان) بر وجود وابستگی مکانی

1. DeMars

2. log likelihood ratio test

3. akaike information criterion (AIC)

4. bayesian information criterion (BIC)

5. Duan, Hardle & Gentle

بین سؤال‌های متن‌ها به کار برده می‌شود. اگر واریانس تصادفی اثر متن صفر باشد، استقلال مکانی سؤال در بین سؤال‌های متن وجود دارد اما اگر صفر نباشد، وابستگی مکانی سؤال بین سؤال‌های یک متن را نشان می‌دهد. طبق مطالعات شبیه‌سازی اگر واریانس تصادفی اثر متن کوچک‌تر از  $0/25$  باشد نشان‌دهنده وابستگی مکانی سؤال ناچیز در بین سؤال‌های یک متن و بالاتر از  $0/25$  نشان‌دهنده وابستگی مکانی قابل ملاحظه سؤال بین سؤال‌های متن است (ژانگ، شن و کندی<sup>۱</sup>، ۲۰۱۰؛ به نقل از روند، ۲۰۱۵). بر طبق مطالعات تجربی نیز اگر واریانس تصادفی اثر متن بین  $0/5$  تا  $2$  باشد اثر قابل ملاحظه محسوب می‌شود. بزرگ بودن اثر متن (آزمونچه)، وجود وابستگی مکانی بین سؤال‌های درون متن را نشان می‌دهد (وانگ، برادللو و واینر<sup>۲</sup>، ۲۰۰۲ و بی. ژانگ<sup>۳</sup>، ۲۰۱۰؛ به نقل از روند، ۲۰۱۵). به‌منظور تعیین اینکه تا چه اندازه وابستگی مکانی ایجاد شده در متن حاصل عامل پیش‌بینی کننده (تک‌زبان-دو‌زبان بودن دانش‌آموزان) است، واریانس اثر متن در مدل بدون پیش‌بینی کننده با واریانس اثر متن در مدل با پیش‌بینی کننده (عامل زبان) برحسب روش بریک و راودنبوش<sup>۴</sup> (۱۹۹۲، به نقل از روند، ۲۰۱۵) مورد مقایسه قرار گرفت. این روش به شرح زیر است:

#### (واریانس اثر متن در مدل بدون پیش‌بینی - واریانس اثر متن در مدل با پیش‌بینی)

##### واریانس اثر متن در مدل بدون پیش‌بینی

همچنین مفروضه تأثیر نیز بررسی شد. این شاخص در صورت معنی‌دار بودن نشان می‌دهد دو گروه از همان ابتدا در توانایی خود با یکدیگر متفاوت هستند در نتیجه، نتایج کارکرد افتراقی سؤال باید با احتیاط تفسیر شود. به‌منظور تحلیل بعدیت، اثر متن و تأثیر از نرم‌افزار Mplus<sup>v/4</sup> استفاده شد.

برای تحلیل کارکرد افتراقی سؤال نیز از رویکرد کای و همکاران (۲۰۱۱) استفاده شد. کای و همکاران (۲۰۱۱) از روش نسبت درست‌نمایی نظریه سؤال-پاسخ برای

1. Zhang, Shen & Cannady

2. Wang, Bradlow & Wainer

3. Zhang

4. Bryk & Raudenbush

کارکرد افتراقی سؤال<sup>۱</sup> (IRTLRDIF) تیسن، اشتینبرگ و واینر<sup>۲</sup> (۱۹۹۳) و آزمون والد برای کارکرد افتراقی سؤال توصیف شده توسط لانگر (۲۰۰۸) برای مدل‌سازی دوعاملی استفاده کردند. این رویکرد یک روش مبتنی بر نظریه سؤال-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 بدون کارکرد افتراقی سؤال همه پارامترهای سؤال‌ها در بین گروه‌ها برابر هستند. در مدل مقایسه برای مطالعه کارکرد افتراقی سؤال محدودیت‌های برابری بین گروهی خوشه‌های سؤال نادیده گرفته می‌شود. در این رویکرد به منظور تحلیل کارکرد افتراقی سؤال‌های هر متن، در هر مرحله سؤال‌های یک متن به عنوان سؤال‌های داوطلب به منظور بررسی کارکرد افتراقی سؤال برگزیده می‌شوند و بقیه سؤال‌ها در خوشه‌های محتوایی دیگر یا متن‌ها (آزمونچه‌ها) به عنوان سؤال‌های لانگر به کار برده می‌شوند تا مقیاس متغیرهای نهفته به یکدیگر پیوند داده شوند. در واقع در هر مرحله، سؤال‌های یک متن به عنوان یک صفت نهفته تک‌بعدی با لانگر قرار دادن سؤال‌های متن‌های دیگر بررسی می‌شود. برای بررسی کارکرد افتراقی سؤال، عرض از مبدأ 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 درون گروه و هم بین گروه به‌طور هم‌زمان برابر قرار داده می‌شوند و واریانس‌های عامل خاص برای همه گروه‌ها به‌جز گروه مرجع که فرض می‌شود واریانس آن برابر یک است، آزادانه برآورد می‌شود (کای و همکاران، ۲۰۱۱). در این پژوهش، گروه تک‌زبان‌ها به عنوان گروه مرجع و دوزبان‌ها به عنوان گروه مقایسه بررسی شد و برای بررسی کارکرد افتراقی سؤال از نرم‌افزار IRTLDRIF استفاده شد.

به منظور برآورد پارامترها هم از روش برآورد اطلاعات کامل بیشینه درست‌نمایی<sup>۳</sup> برای داده‌های ناکامل<sup>۴</sup> استفاده شد. چن و فینبرگ<sup>۵</sup> (۱۹۷۴)، به نقل از لینگ‌پینگ و اسلام<sup>۶</sup> (۲۰۰۸) بیان کرده‌اند روش بیشینه درست‌نمایی برای جداول توافقی ناکامل،

1. IRT likelihood ratio DIF (IRTLRDIF)

2. Thissen, Steinberg & Wainer

3. full information maximum marginal likelihood estimation

4. incomplete data

5. Chen & Fienberg

6. Ling Ping & Islam

هنگامی که ردیف‌ها و ستون‌های داده‌ها یا هر دو دارای مقادیر گم‌شده هستند، به کار برده می‌شود. به منظور استفاده از این روش از الگوریتم انتظار-بیشینه‌سازی<sup>۱</sup> 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.

### یافته‌ها

#### الف- بررسی انحراف از تک‌بعدی بودن

جدول (۲) نتایج بررسی وجود انحراف از تک‌بعدی بودن در متن‌های خواندن مطالعه پرلز ۲۰۱۱ ایران بر مبنای مدل پاسخ مدرج شده و مدل دوعاملی نظریه سؤال-پاسخ را نشان می‌دهد.

جدول (۲) شاخص‌های برازندگی مدل‌های پاسخ مدرج شده و دوعاملی نظریه سؤال-پاسخ در

متن‌های پرلز ۲۰۱۱ ایران

معناداری	تبدیل تفاوت $Z_{-2II}$ به $Z$	پارامتر دو	تفاوت نسبت درست‌نمایی دو مدل	تعداد پارامتر	دوعاملی	تعداد پارامتر	مدل پاسخ مدرج شده	شاخص
<۰/۰۱	۹۶/۳۶	۱۴۶	-۴۶۴۳/۳۶۶	۸۰۶	۱۳۴۸۶۵/۸۸	۶۶۰	۱۳۹۵۰۹/۲۴۶	لگاریتم نسبت درست‌نمایی (-2II)
-	-	-	-	-	۲۷۱۳۴۳/۷۶۱	-	۲۸۰۳۳۸/۴۹۲	شاخص AIC
-	-	-	-	-	۲۷۷۲۱۵/۸۳۲	-	۲۸۳۶۲۷/۱۸۲	شاخص BIC

نتایج جدول (۲) نشان می‌دهد شاخص‌های برازندگی در مدل پاسخ مدرج شده و مدل دوعاملی نظریه سؤال-پاسخ (AIC=۲۸۰۳۳۸/۴۹۲، BIC=۲۸۳۶۲۷/۱۸۲،  $p < ۰/۰۱$ ،  $-2II(660) = ۱۳۹۵۰۹/۲۴۶$ ) و (AIC=۲۷۱۳۴۳/۷۶۱، BIC=۲۷۷۲۱۵/۸۳۲،  $p < ۰/۰۱$ ،  $-2II(806) = ۱۳۴۸۶۵/۸۸$ ) هستند. مقایسه لگاریتم نسبت درست‌نمایی دو

<sup>۱</sup>. expectation-maximum algorithm

مدل نشان داد لگاریتم نسبت درست‌نمایی مدل دوعاملی کوچک‌تر و برازندگی بهتر را نشان می‌دهد. نتایج مقایسه تفاوت لگاریتم نسبت درست‌نمایی دو مدل نیز معنی‌دار است ( $0/01 < p < 93/36, z = -643/366, t(146) = -211$ )؛ در نتیجه اثراتی از اثر متن یا وابستگی مکانی سؤال در متن‌های خواندن مطالعه پرلز وجود دارد. نتایج مقایسه شاخص ملاک اطلاعات آکائیک و ملاک اطلاعات بیزی نیز نشان داد این شاخص‌ها در مدل دوعاملی نسبت به مدل پاسخ مدرج شده کوچک‌تر هستند در نتیجه مدل دوعاملی نظریه سؤال-پاسخ بر ارزش بهتری با داده‌ها نسبت به مدل پاسخ مدرج شده دارد.

### ب- اثر متن یا وابستگی مکانی سؤال

اثر متن‌های پرلز در جدول (۳) ارائه شده است. واریانس اثر متن (آزمونچه) نشان‌دهنده درجه استقلال مکانی سؤال در بین سؤال‌های یک متن است.

جدول (۳) نتایج واریانس‌های تصادفی اثر متن‌ها (آزمونچه) در مطالعه پرلز ایران

وابستگی مکانی سؤال		خطای استاندارد	واریانس اثر متن	صفات نهفته
مطالعات تجربی	مطالعات شبیه‌سازی			
ملاک ۰/۵ تا ۲	ملاک ۰/۲۵			
ناچیز	ناچیز	۰/۰۱۵	۰/۰۹۸	درک مطلب کلی
ناچیز	زیاد	۰/۱۰۴	۰/۴۰۷**	درک مطلب ادبی ۱
ناچیز	زیاد	۰/۰۷۸	۰/۳۵۷**	درک مطلب ادبی ۲
ناچیز	ناچیز	۰/۰۲۱	۰/۰۵۴	درک مطلب ادبی ۳
ناچیز	ناچیز	۰/۰۲۶	۰/۱۲۹	درک مطلب ادبی ۴
ناچیز	ناچیز	۰/۰۱۸	۰/۰۴	درک مطلب ادبی ۵
ناچیز	ناچیز	۰/۰۱۲	۰/۰۲	درک مطلب اطلاعاتی ۱
ناچیز	ناچیز	۰/۰۱۴	۰/۰۳۹	درک مطلب اطلاعاتی ۲
ناچیز	ناچیز	۰/۰۵۳	۰/۲۱۶	درک مطلب اطلاعاتی ۳
ناچیز	ناچیز	۰/۰۴۴	۰/۲۳۷	درک مطلب اطلاعاتی ۴
ناچیز	ناچیز	۰/۰۴۵	۰/۲۳۲	درک مطلب اطلاعاتی ۵

نتایج جدول (۳) نشان می‌دهد در مطالعه پرلز ۲۰۱۱ در صفات نهفته خاص مرتبط با دو متن ادبی (صفات درک مطلب ادبی 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 نشان‌دهنده 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 سؤال‌ها را نشان می‌دهد.

### ج- اثر زبان بر وابستگی مکانی بین سؤال‌ها

جدول (۴) نتایج درصد واریانس تبیین شده متن‌های دارای وابستگی مکانی سؤال توسط عامل تفاوت زبانی را نشان می‌دهد.

جدول (۴) نتایج تحلیل تأثیر زبان بر وابستگی مکانی بین سؤال‌های متن‌ها

متن	واریانس اثر متن	واریانس اثر زبان	واریانس تبیین شده توسط زبان
مهارت درک مطلب ادبی اول	۰/۴۰۷	۰/۰۹۲	۰/۷۷۴
مهارت درک مطلب ادبی دوم	۰/۳۵۷	۰/۰۶۶	۰/۸۱۵

نتایج جدول (۴) نشان می‌دهد که زبان در متن اول مرتبط با درک مطلب ادبی ۷۷/۴ درصد واریانس و در متن دوم مرتبط با درک مطلب ادبی ۸۱/۵ درصد واریانس وجود وابستگی مکانی بین سؤال‌های متن‌ها را تبیین می‌کند. این نتایج نشان می‌دهد این دو متن تا حد زیادی در بین دو گروه دانش‌آموزان تک‌زبان-دوزبان به گونه متفاوت ادراک می‌شوند و تا حد زیادی تحت تأثیر زبان هستند.

### د- تحلیل شاخص تأثیر

جدول (۵) نتایج بررسی تأثیر زبان (تک‌زبان-دوزبان بودن) به‌عنوان یک پیش‌بینی‌کننده سطح فرد بر تفاوت در برآورد توانایی‌های دانش‌آموزان دو گروه یا شاخص تأثیر را نشان می‌دهد.

جدول (۵) شاخص تأثیر زبان

خطای استاندارد	شاخص تأثیر	تک‌زبان-دوزبان بودن
۰/۰۶۲	-۰/۲۲۲**	

نتایج جدول (۵) نشان می‌دهد اثر زبان (شاخص تأثیر) به‌طور معنی‌داری متفاوت از صفر است (-۰/۲۲۲)، علامت منفی آن نشان می‌دهد تک‌زبان‌ها (به‌عنوان گروه مرجع) به‌طور متوسط برآوردهای توانایی معنی‌دار بالاتری نسبت به دوزبان‌ها در متن‌های خواندن پرلز دارند.

#### ه- کارکرد افتراقی سؤال متن‌های پرلز

جدول (۶) نتایج تحلیل کارکرد افتراقی سؤال متن‌های مطالعه پرلز را نشان می‌دهد (به علت تعداد زیاد جداول، نتایج متن (آزمونچه) اول به‌عنوان نمونه آورده شده است). نتایج جدول (۶) نشان می‌دهد بین دانش‌آموزان تک‌زبان و دوزبان، ۲۸ سؤال (۱۹/۱۸ درصد) دارای کارکرد افتراقی یکنواخت (هماهنگ)، ۲۷ سؤال (۱۸/۴۹ درصد) دارای کارکرد افتراقی غیر یکنواخت (غیر هماهنگ) و ۹۱ سؤال (۶۲/۳۳ درصد) هم در دو گروه معادل (هم‌ارزی سؤال) هستند. از بین ۲۸ سؤال دارای کارکرد افتراقی سؤال یکنواخت ۳ سؤال چهارگزینه‌ای (۲/۰۵ درصد) و ۲۵ سؤال پاسخ‌ساز (۱۷/۱۳ درصد) و از بین ۲۷ سؤال دارای کارکرد افتراقی سؤال غیریکنواخت ۱۴ سؤال چهارگزینه‌ای (۹/۵۹ درصد) و ۱۳ سؤال پاسخ‌ساز (۸/۹ درصد) هستند. همچنین از بین ۹۱ سؤال هم‌ارز ۵۱ سؤال چهارگزینه‌ای (۳۴/۹۳ درصد) و ۴۰ سؤال پاسخ‌ساز (۲۷/۴ درصد) هستند. نتایج نشان داد در سؤال‌های دارای کارکرد افتراقی یکنواخت (هماهنگ) شیب سؤال‌ها در بین دو گروه دانش‌آموز تک‌زبان و دوزبان یکسان است و فقط آستانه بین طبقات سؤال‌ها در بین دو گروه در پیوستار توانایی با یکدیگر متفاوت است. بعضی از سؤال‌ها برای دانش‌آموزان تک‌زبان و بعضی دیگر برای دانش‌آموزان دوزبان به تناوب، آستانه دشواری بالاتری دارند. در سؤال‌های دارای کارکرد افتراقی غیریکنواخت (ناهماهنگ) نیز سؤال‌ها به تناوب دارای شیب و آستانه دشواری بیشتر و کمتر در هر دو گروه دانش‌آموزان تک‌زبان و دوزبان هستند. در همه سؤال‌ها نیز میانگین دانش‌آموزان تک‌زبان بالاتر از دانش‌آموزان دوزبان است



و سطح پیشرفت دانش‌آموزان تک‌زبان از همگونی نسبی بیشتری نسبت به دانش‌آموزان دوزبان برخوردار است. در بقیه سؤال‌ها بین آستانه دشواری و شیب سؤال‌ها در دو گروه تفاوت معنی‌داری در پیوستار توانایی وجود ندارد.

جدول (۶) نتایج کارکرد افتراقی سؤال‌های متن اول پرلز ۲۰۱۱ بین دو گروه تک‌زبان-دوزبان

DIF	گروه کانونی (دوزبان)				گروه مرجع (تک‌زبان)				فرض برابری همه پارامترها			سؤال
	انحراف استاندارد	میانگین	آستانه دوم	آستانه اول	شیب	آستانه دوم	آستانه اول	شیب	وضعیت فرض	درجه آزادی	مقدار $\chi^2$	
خیر	۱/۰۷	-۰/۵۴	-	-۰/۸۶	۱/۲۵	-	-۱/۰۹	۱/۲۲	قبول	۲	۳/۲	۱
خیر	۱/۰۷	-۰/۵۴	-	-۰/۴۹	۰/۷۹	-	-۰/۶۸	۰/۷	قبول	۲	۱	۲
خیر	۱/۰۷	-۰/۵۴	-	-۱/۱۷	۱/۵۴	-	-۱/۴۴	۱/۳۱	قبول	۲	۲/۷	۳
خیر	۱/۰۷	-۰/۵۴	-	-۱/۳۵	۱/۰۸	-	-۱/۵۱	۱/۰۸	قبول	۲	۱/۲	۴
خیر	۱/۰۷	-۰/۵۴	-	-۱/۵۳	۰/۸۲	-	-۱/۷۱	۰/۷۴	قبول	۲	۰/۴	۵
غیریکنواخت	۱/۰۷	-۰/۵۴	-	-۰/۲۷	۰/۹۹	-	-۰/۵۹	۰/۶۸	رد	۲	۵/۷	۶
یکنواخت	۱/۰۷	-۰/۵۴	۱/۶۳	-۱/۱۳	۰/۷۹	۱/۴۲	۰/۷۸	۰/۸۹	رد	۳	۴/۵	۷
غیریکنواخت	۱/۰۷	-۰/۵۴	-	-۰/۳۴	۱/۶۴	-	-۰/۵۶	۱/۳۱	رد	۲	۴/۹	۸
غیریکنواخت	۱/۰۷	-۰/۵۴	۰/۱۴	-۰/۶۴	۲/۰۸	۰/۰۹	-۰/۸۳	۱/۲۸	رد	۳	۱۴/۱	۹
یکنواخت	۱/۰۷	-۰/۵۴	-	-۱/۱۲	۰/۹۸	-	-۱/۵۱	۰/۷۹	قبول	۲	۲/۶	۱۰
خیر	۱/۰۷	-۰/۵۴	-	-۰/۴۴	۰/۵۸	-	-۰/۵۴	۰/۴۶	قبول	۲	۰/۶	۱۱
یکنواخت	۱/۰۷	-۰/۵۴	۲/۵۴	۱/۰۳	۱/۲۱	۲/۲۳	۰/۷۹	۱/۱۵	رد	۳	۵/۱	۱۲
خیر	۱/۰۷	-۰/۵۴	-	-۰/۶۸	۰/۷۴	-	-۰/۸۵	۰/۶۹	قبول	۲	۰/۶	۱۳

### بحث و نتیجه‌گیری

این مطالعه با هدف بررسی اصل همگونی زبانی یا یکسان بودن سطح زبان مورد نیاز برای متن‌های خواندن مطالعه پرلز ۲۰۱۱ در بین دانش‌آموزان تک‌زبان-دوزبان ایرانی انجام گرفت. به این منظور، بعدیت یا انحراف از تک‌بعدی بودن، وابستگی مکانی

سؤال، تأثیر یا تفاوت در برآورد توانایی‌های دانش‌آموزان و کارکرد افتراقی سؤال متن‌های پرلز ۲۰۱۱ ایران برحسب زبان (تک‌زبان-دو‌زبان بودن) دانش‌آموزان تحلیل شد. مشکل اصلی آزمون‌های مبتنی بر متن (آزمونچه)، تخطی از مفروضه استقلال مکانی سؤال هنگام کاربرد مدل‌های تک‌بعدي نظریه سؤال-پاسخ به واسطه وابستگی درونی سؤال‌های درون هر متن (آزمونچه) است؛ در این حالت استفاده از مدل‌های تک‌بعدي موجب برآزش مناسب با داده‌ها نمی‌شود و این امر می‌تواند در دقت برآورد پارامترهای سؤال و توانایی تأثیر بگذارد. به‌منظور بررسی بعدیت، وابستگی مکانی سؤال و کارکرد افتراقی سؤال آزمون‌های مبتنی بر متن، امروزه مدل‌های مبتنی بر متن (آزمونچه) همچون مدل دو‌عاملی نظریه سؤال-پاسخ کاربرد زیادی در سنجش‌های آموزشی و روان‌شناختی پیدا کرده است.

نتایج بررسی انحراف از تک‌بعدي بودن صفات نهفته زیربنایی سؤال‌های متن‌های خواندن پرلز نشان داد بین شاخص‌های برازندگی دو مدل پاسخ مدرج شده و مدل دو‌عاملی، تفاوت معنی‌داری وجود دارد و مدل دو‌عاملی نظریه سؤال-پاسخ برازندگی بهتری با سؤال‌های متن‌ها دارد. نتایج شیب‌های سؤال در مدل پاسخ مدرج شده نیز نشان داد همه سؤال‌ها دارای شیب مثبت اما کوچک (زیر ۰/۸) روی صفت نهفته (درک مطلب) هستند. نتایج بررسی شیب‌ها در مدل دو‌عاملی نشان داد شیب سؤال‌ها روی صفات نهفته خاص (درک مطلب ادبی، درک مطلب اطلاعاتی) بزرگ‌تر از صفت نهفته کلی (درک مطلب) است. این نتایج نشان می‌دهد که سؤال‌ها با صفت نهفته خود همبستگی بالاتری نسبت به صفت نهفته کلی دارند. دی‌مارس (۲۰۱۳) در تفسیر شیب‌های سؤال‌ها در مدل دو‌عاملی بیان می‌کند چنانچه شیب‌ها در دامنه ۰ تا ۰/۸ قرار گیرند 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 تک‌بعدي باشد، شیب بیشتر سؤال‌ها روی صفت نهفته زیربنایی باید بزرگ و در دامنه ۰/۸ تا ۱/۴۹ و یا ۱/۵ به بالا قرار گیرند. با توجه به شاخص‌های برازندگی و شیب سؤال‌ها در دو مدل، اثراتی از انحراف از تک‌بعدي بودن در سؤال‌های متن‌های پرلز ۲۰۱۱ ایران مشاهده شد. همان‌طور که دی‌مارس (۲۰۱۳) بیان می‌کند چون در مدل دو‌عاملی صفات نهفته خاص متعامد با صفت کلی هستند ماتریس باقی‌مانده، واریانس مشترک بین زیرمجموعه‌ای از سؤال‌ها را تبیین می‌کند و این واریانس فراتر از واریانس تبیین شده توسط صفت کلی است. بنابراین شیب بزرگ سؤال‌های هر خوشه روی صفت نهفته

خاص خود و افزایش مقدار واریانس تبیین شده هر سؤال با اضافه شدن صفات نهفته خاص به مدل نشان می‌دهد که علاوه بر صفت کلی، صفات نهفته خاص نیز در تبیین واریانس سؤال‌ها نقش دارند. این نتایج، نشان‌دهنده وجود عوامل ثانوی است که بر عملکرد دانش‌آموزان تک‌زبان و دوزبانه ایرانی تأثیر دارند. به گفته دی‌مارس (۲۰۰۶) این عوامل می‌توانند داشتن دانش قبلی در ارتباط با موضوع متن، داشتن یک مهارت خاص، سطح علاقه و یا عوامل انگیزشی خاص برای یادگیرندگان باشند.

نتایج بررسی اثر متن نشان داد وابستگی مکانی سؤال در دو صفت نهفته (درک مطلب ادبی) مرتبط با متن‌های گل‌های روی پشت بام و پرواز کن عقاب، پرواز کن موجب انحراف از تک‌بعدی بودن سؤال‌های متن‌های پرلز شده است زیرا واریانس تصادفی این دو متن بزرگ‌تر از ۰/۲۵ هستند که برحسب نتایج مطالعات شبیه‌سازی دارای وابستگی مکانی قابل ملاحظه سؤال‌ها هستند. این نتایج، وجود عوامل تأثیرگذار دیگری را بر عملکرد دانش‌آموزان در این دو متن نشان می‌دهد. نتایج بررسی تأثیر زبان (تک‌زبان- دوزبان) دانش‌آموزان بر ایجاد وابستگی مکانی بین سؤال‌های دو متن نشان داد در صفت نهفته خاص مرتبط با متن گل‌های روی پشت بام ۷۷/۴ و صفت نهفته خاص مرتبط با متن پرواز کن عقاب، پرواز کن ۸۱/۵ درصد از وابستگی مکانی بین سؤال‌ها را تفاوت زبانی دانش‌آموزان تبیین می‌کند. این نتایج نشان می‌دهد تفاوت زبانی دانش‌آموزان تأثیر زیادی بر عملکرد دانش‌آموزان در این دو شایستگی دارد. به طور کلی نتایج نشان داد در این دو متن علاوه بر صفت نهفته خاص اندازه‌گیری شده توسط هر یک از متن‌ها، تفاوت زبانی یکی از پیش‌بینی‌کننده‌های مهم تفاوت در عملکرد دانش‌آموزان تک‌زبان-دوزبان در این دو متن است به این معنا که زبان اثر یکسانی بر همه دانش‌آموزان و همه سؤال‌های این دو متن ندارد. نتایج مقایسه واریانس تصادفی اثر متن‌ها بین تک‌زبان‌ها و دوزبان‌ها در متن گل‌های روی پشت بام نشان داد واریانس اثر متن در گروه تک‌زبان‌ها و دوزبان‌ها به ترتیب ۰/۱۰۷ و ۰/۱۹۴ و در متن پرواز کن عقاب، پرواز کن به ترتیب ۰/۱۳۶ و ۰/۲۶۴ است. با توجه به این نتایج می‌توان گفت وابستگی مکانی سؤال در بین دوزبان‌ها بیشتر از تک‌زبان‌ها به خصوص در متن پرواز کن عقاب، پرواز کن است و زبان علاوه بر صفت خاص مرتبط با متن اندازه‌گیری شده توسط سؤال‌های متن، تأثیر زیادی بر عملکرد دانش‌آموزان دوزبان نسبت به دانش‌آموزان تک‌زبان دارد.

عوامل زمینه‌ای متعددی وجود دارند که می‌توانند در بین گروه‌های مختلف موجب وابستگی مکانی سؤال شوند. تائو (۲۰۰۸) اظهار داشته که بر اساس نتایج پژوهش‌های لی<sup>۱</sup> (۲۰۰۴)؛ سیرسی، تیسن و واینر<sup>۲</sup> (۱۹۹۱)؛ تیسن، اشتینبرگ و مونی<sup>۳</sup> (۱۹۸۹) و ین (۱۹۹۳)، مهم‌ترین این عوامل در مطالعات بین‌المللی از جمله پرلز، عامل وابستگی متن<sup>۴</sup> و در مرحله بعد وابستگی شکل سؤال یا پاسخ<sup>۵</sup> هستند. وابستگی متن بیان می‌کند که داشتن اطلاعات بیشتر یا علاقه بیشتر به موضوع متن در بین بعضی از دانش‌آموزان می‌تواند احتمال پاسخ درست دادن به سؤال‌ها را افزایش دهد. نتایج نشان داد در متن‌های گل‌های روی پشت بام و پرواز کن عقاب پرواز کن، پراکندگی در داشتن اطلاعات درباره موضوع متن و یا سطح علاقه در بین دوزبان‌ها بیشتر از تک‌زبان‌ها است. عامل زمینه‌ای دیگری که احتمالاً موجب وابستگی مکانی سؤال شده است، وابستگی شکل سؤال یا پاسخ است. این نوع وابستگی زمانی روی می‌دهد که بعضی از دانش‌آموزان در نوشتن پاسخ درست، داستان و یا ترسیم یک تصویر بهتر از دانش‌آموزان دیگر عمل می‌کنند. این موجب دادن پاسخ‌های متنوع به سؤال‌ها برحسب طول و نوع متفاوت پاسخ مورد نیاز سؤال‌ها مثل پاسخ‌ساز و یا چندگزینه‌ای می‌شود. این مورد به وضوح در پرلز وجود دارد زیرا مطالعه پرلز یک مطالعه مبتنی بر متن است که متن‌های آن دارای سؤال با شکل آمیخته هستند. نتایج مقایسه فراوانی‌های پاسخ به سؤال‌های چندگزینه‌ای و کوتاه‌پاسخ نشان داد در ایران در بین تک‌زبان‌ها و دوزبان‌ها تفاوت محسوسی وجود ندارد اما در سؤال‌های پاسخ‌ساز نمره‌گذاری شده به صورت سه یا چهارارزشی، تک‌زبان‌ها نسبت به دوزبان‌ها فراوانی تعداد پاسخ بیشتری دارند. این نتایج نشان می‌دهد تا حدودی در مطالعه پرلز اثر شکل سؤال وجود دارد. هنگامی اثر شکل سؤال روی می‌دهد اضافه کردن سؤال‌های پاسخ‌ساز گستره پاسخ می‌تواند به پیچیدگی آزمون بینجامد در این حالت به احتمال زیاد سؤال‌های چندگزینه‌ای و پاسخ‌ساز توانایی‌های متفاوتی را اندازه‌گیری می‌کنند که می‌تواند به بعدیت منجر شود.

---

1. Lee

2. Sireci, Thissen & Wainer

3. Thissen, Steinberg & Mooney

4. passage dependence

5. item or response format

با این وجود، همان‌طور که زینیسکی، هامبلتون و سیرسی<sup>۱</sup> (۲۰۰۳) بیان کرده‌اند اگرچه وابستگی مکانی سؤال نامطلوب است اما دلایل خوبی وجود دارد سؤال‌هایی که وابستگی درونی دارند در سنجش به کار برده شوند زیرا بسیاری از تکالیف در دنیای واقعی نیازمند مسائل مبتنی بر حل کردن یا حل یک مشکل در حالت گام به گام هستند، بنابراین سؤال‌های وابسته به زمینه در یک آزمون ممکن است روایی سازه را افزایش دهد. از این‌رو، چالش توسعه‌دهندگان آزمون، حذف وابستگی سؤال نیست بلکه این است چگونه به‌درستی چنین وابستگی‌هایی را مدل‌سازی کنند که وابستگی مکانی سؤال روی ندهد.

نتایج بررسی تفاوت در برآورد توانایی‌های فرد به‌عنوان تابعی از زبان (شاخص تأثیر) نیز نشان داد به‌طور متوسط برآوردهای توانایی دانش‌آموزان تک‌زبانه نسبت به دانش‌آموزان دوزبانه بالاتر است. این نشان‌دهنده عملکرد افتراقی در توانایی دو گروه است که از تفاوت در متوسط توانایی‌های دو گروه به وجود می‌آید و نشان می‌دهد دو گروه در سطوح توانایی‌هایشان برحسب عملکرد در سؤال با یکدیگر متفاوت هستند. در نتیجه، هنگام بررسی کارکرد افتراقی سؤال‌ها باید با احتیاط عمل کرد و نمی‌توان این تفاوت را دقیقاً به تفاوت عملکرد سؤال در دو گروه در سطوح یکسانی از توانایی نسبت داد؛ زیرا ممکن است به تفاوت واقعی در توانایی دو گروه در متن‌های خواندن ارتباط داشته باشد. البته باید به این موضوع توجه کرد که در طی سه دوره برگزار شده مطالعه پرلز، نتایج نشان داده است دوزبانه‌ها عملکرد پایین‌تری نسبت به تک‌زبانه‌ها در متن‌ها دارند، در حالی است که در مطالعات دوزبانگی به‌خصوص در آمریکا نتایج نشان داده است که دوزبانه‌ها در بعضی مهارت‌های رشد شناختی همچون کنترل اجرایی، آگاهی فرازبانی، حافظه فعال، آگاهی فراشناختی و تفکر واگرا در سطح بالاتری نسبت به تک‌زبانه‌ها قرار دارند (مورتن و هارپر، ۲۰۰۷ به نقل از سیاهابودین، ۲۰۱۳).

نتایج مقایسه دانش‌آموزان تک‌زبانه و دوزبانه ایرانی نشان داد در سؤال‌ها با کارکرد افتراقی یکنواخت، در سؤال‌های دوازده‌گانه (۵ به ۱) و در سؤال‌های چهارده‌گانه (۱۵ به ۷) برای تک‌زبانه‌ها نسبت به دوزبانه‌ها دشوارتر هستند. در سؤال‌ها با کارکرد افتراقی غیریکنواخت نیز در سؤال‌های دوازده‌گانه (۱۲ به ۱) و در سؤال‌های

1. Zenisky, Hambleton & Sireci

چندارزشی (۷ به ۷) سؤال‌های دارای شیب بزرگ‌تر برای تک‌زبانان نسبت به دوزبانان هستند. نتایج نشان داد سؤال‌های دوارزشی و چندارزشی با کارکرد افتراقی یکنواخت برای دوزبانان نسبت به تک‌زبانان آسان‌تر هستند اما در سؤال‌های دوارزشی دارای کارکرد افتراقی غیریکنواخت با افزایش توانایی سرعت احتمال پاسخ درست در بین تک‌زبانان نسبت به دوزبانان با احتمال بیشتری افزایش پیدا می‌کند.

این نتایج نشان می‌دهد زمانی که شیب سؤال‌ها در بین دو گروه یکسان است، دوزبانان در مقایسه با تک‌زبانان به واسطه مهارت‌هایی همچون تغییر بین وظایف در آمیج‌های ذهنی، به‌روز کردن و نظارت بر بازنمایی‌های حافظه، بازداری پاسخ‌های بارز، تفکر کردن درباره زبان برحسب شکل‌ها و ساختار و رابطه بین دو زبان برای تولید معانی جامع، ذخیره موقتی اطلاعات به کار برده شده برای عملکرد در انواع متنوع تکالیف، آگاهی فرد از دانش خود درباره فعالیت‌های ذهنی و راهبردهای یادگیری و فرایندهای ذهنی و روش‌های برانگیختن خلاقیت با نتایج نوعی متنوع، عملکرد بهتری در سؤال‌ها دارند و سؤال‌ها برای آنها آسان‌تر از تک‌زبانان هستند. اما زمانی که احتمال پاسخ درست با افزایش توانایی به‌سرعت افزایش پیدا می‌کند به‌خصوص زمانی که مهارت‌های شناختی سطح بالا برای پاسخ دادن به سؤال‌ها نیاز نیست (مهارت بازشناسی در سؤال‌های چندگزینه‌ای) تک‌زبانان نسبت به دوزبانان احتمال پاسخ درست بیشتری دارند و از مهارت بازشناسی بهتر استفاده می‌کنند.

اما اینکه چرا دوزبانان در کل آزمون عملکرد ضعیف‌تری نسبت به تک‌زبانان دارند را شاید بتوان از دیدگاه بیکر<sup>۱</sup> (۲۰۱۱) تبیین کرد. به گفته بیکر (۲۰۱۱) عواملی همچون توانایی به‌کارگیری فعال زبان دوم، کاربرد زبان دوم در موقعیت‌های متنوع، توانایی برقراری تعادل و به‌کارگیری دو زبان، سن فرد، دوره تحول (رشد) فرد، یادگیری فرهنگ زبان دوم توأم با یادگیری زبان دوم، بافت و محیط زندگی افراد و ترجیح زبان دوم بر زبان مادری در کلاس درس می‌توانند بر عملکرد دوزبانان تأثیر بگذارد. یکی اینکه دوزبانان ایرانی، دوزبانان هستند که از همان ابتدا در ایران متولد می‌شوند و همانند دوزبانان کشورهای آمریکایی که بیشترین تفاوت بین دوزبانان و تک‌زبانان را در مطالعات خواندن نشان داده‌اند از کشورهایی همچون آمریکای جنوبی، آفریقا یا کشورهای آسیایی به این کشور مهاجرت نکرده‌اند. در

<sup>۱</sup>. Baker

کشورهایی همانند آمریکا دوزبانه‌ها مجبورند برای برآوردن نیازهای خود در محیط بیرون از خانه از زبان غالب جامعه در کنار زبان مادری استفاده زیادی کنند یعنی به‌کارگیری زبان دوم در موقعیت‌های متنوع و این می‌تواند موجب افزایش توانایی به‌کارگیری فعال زبان دوم و برقراری تعادل بین دو زبان به‌منظور برآوردن نیازهای خود شود. نتیجه آن می‌تواند رشد مهارت‌های شناختی به واسطه استفاده هم‌زمان از دو زبان باشد در حالی که در ایران دوزبانه‌ها بیشتر در استان‌های مرزی ایران زندگی می‌کنند و افراد زبان مادری را بیشتر از زبان رسمی جامعه برای رفع نیازهای خود در محیط خانه و جامعه به کار می‌برند؛ این می‌تواند مانع رشد مطلوب مهارت‌های شناختی ذکر شده در دوزبانه‌های ایرانی شود. علت دوم را شاید بتوان به بافت و محیط زندگی افراد و ترجیح زبان دوم بر زبان اول در کلاس درس نسبت داد. از آنجایی که معلمان استان‌های دوزبانه معمولاً بومی هستند و همان زبانی را در کلاس درس به کار می‌گیرند که زبان مادری دانش‌آموزان دوزبانه محسوب می‌شود. اگر معلمان به‌خصوص در سال‌های اولیه ابتدایی، زبان فارسی را کمتر از زبان مادری دانش‌آموزان در کلاس درس به کار گیرند، می‌تواند رشد مهارت‌های شناختی دوزبانه‌ها را با مشکل روبه‌رو کنند. علت سوم را شاید بتوان در محتواها و شیوه‌های تدریس زبان دوم به دوزبانه‌ها جستجو کرد. به نظر می‌رسد با توجه به عملکرد ایران در سه دوره برگزاری پرلز و عملکرد ضعیف دانش‌آموزان ایرانی، برنامه درسی *بنخوانیم و بنویسیم* ایران هنوز به لحاظ هدف، محتوا، روش‌های تدریس و روش‌های ارزشیابی به اصلاحات اساسی نیاز دارد. علت چهارم را شاید بتوان از نظر ویی<sup>۱</sup> (۲۰۱۰، به نقل از سیاهابودین، ۲۰۱۵) بیان کرد. به گفته وی شاید علت این باشد که ذهن بچه فقط می‌تواند در یک زمان یکسان فقط بر یک زبان تسلط پیدا کند یعنی تسلط بر بیش از یک زبان دشوار است. این ایده از سوی کسانی مطرح شده است که با نقایصی در اندازه دایره لغات کسب شده، کفایت و عملکرد در دو زبان و معنا و تصویرسازی دوزبانه‌ها روبرو شده‌اند.

به‌طور کلی، توسعه‌دهندگان آزمون‌های بین‌المللی برای اندازه‌گیری سازه مورد نظر خود تلاش می‌کنند عوامل بالقوه‌ای را که به وسیله ویژگی‌های غیر مرتبط با سازه همچون ویژگی‌های زبانی، ارتباطی، شناختی، فرهنگی و فیزیکی یا دیگر ویژگی‌ها

<sup>۱</sup>. Wei

اعمال می‌شوند، کنترل کنند. این مطالعات معمولاً مطالعات چندزبانی هستند و مفروضه ضمنی آنها این است که ترجمه دقیق ابزار می‌تواند به انتقال کامل همه ویژگی‌های اندازه‌گیری همانند روایی و پایایی سازه منجر شود. انتقال کامل ویژگی‌های ابزار، معرف رهایی از سوگیری و اکتساب بالاترین سطح ممکن معادل بودن (هم‌ارزی) است. امروزه برای تحلیل هم‌ارزی‌های مختلف آزمون‌های بین‌المللی همچون پرلز که زبان در آنها نقش مهمی ایفا می‌کند، مدل‌های جدیدی مطرح شده است. مدل دوعاملی نظریه سؤال-پاسخ از جمله مهم‌ترین مدل‌هایی است که در سنجش‌های آموزشی و روان‌شناختی مبتنی بر متن به کار برده می‌شود. این مدل به منظور تحلیل اثرات وابستگی‌های اضافی اعمال شده به واسطه کاربرد متن‌ها در سؤال‌های آزمون‌های استاندارد شده همچون پرلز، پیزا و تیمز به کار برده می‌شود. همچنین مدل‌هایی همچون مدل چندگروهی دوعاملی نظریه سؤال-پاسخ به منظور آشکارسازی کارکرد افتراقی سؤال آزمون‌های مبتنی بر متن که وابستگی مکانی بین سؤال‌ها را به واسطه استفاده از متن‌ها در هنگام تحلیل کارکرد افتراقی سؤال‌ها نسبت به مدل‌های سنتی نظریه سؤال-پاسخ در نظر می‌گیرد، تدوین شده است. 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 پارامترهای سؤال‌ها، دچار سوگیری همانند آنچه در مدل‌های سنتی نظریه سؤال-پاسخ روی می‌دهد، 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؛ زیرا هنگام تحلیل کارکرد افتراقی سؤال نیاز است طبقات جامعه مورد مقایسه قرار گیرد و این می‌تواند کارآمدی این روش‌ها را در برآورد پارامترهای سؤال به علت طبقه‌بندی کردن نمونه کلی به زیر طبقات و کاهش تعداد افراد درون طبقات، کاهش دهد. دوم اینکه روش‌های متعددی برای بررسی بعدیت و کارکرد افتراقی سؤال وجود دارد که بیشتر با داده‌های شبیه‌سازی شده انجام گرفته است. برای داده‌های واقعی مطالعات بین‌المللی به صورت گزینش از بین سؤال‌ها و بیشتر روی سؤال‌های دوازده‌گانه کار شده است و خیلی کم به سؤال‌های ترکیبی چندگزینه‌ای و چندارزشی توجه شده است. همچنین بیشتر مدل‌ها برازش تقریبی را نشان می‌دهند و ممکن است مدل‌های مختلفی با داده‌ها برازش داشته باشد. از جمله مطالعاتی که می‌توان در این



زمینه برای نتیجه‌گیری دقیق‌تر انجام داد، یک مطالعه طولی روی چهار دوره برگزار شده (نتایج مطالعه ۲۰۱۶ پرلز هنوز منتشر نشده است) به‌منظور بررسی روند وجود اثر متن بر بعدیت و وابستگی مکانی سؤال به‌خصوص روی متن‌های مشترک به‌کاربرده شده در این چهار دوره است. همچنین از آنجایی که وجود کارکرد افتراقی سؤال در گروه‌های مختلف می‌تواند تحت تأثیر عوامل مختلفی از جمله استدلال افتراقی، فعالیت‌های آموزشی و شکل آزمون باشد، بهتر است در مطالعات دیگر اثر این عوامل بر آزمون‌های درک مطلب در دانش‌آموزان بررسی شود.

## 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 & Behavioral Sciences*, 46, 135 – 140. <https://doi:10.1016/j.sbspro.2012.05.082>
- Baker, C. (2011). *Foundations of bilingual education and bilingualism* (5th ed.). Bristol: Multilingual Matters.
- Beretvas, S. N. & Walker, C. M. (2012). Distinguishing differential testlet functioning from differential bundle functioning using the multilevel measurement model. *Educational & Psychological Measurement*, 2 (2), 200–223. <https://doi:10.1177/0013164411412768>.
- Beretvas, S. N.; Cawthon, S. W.; Lockhart, L. L. & Kaye, A. D. (2012). Assessing Impact, DIF, and DFF in Accommodated Item Scores: A Comparison of Multilevel Measurement Model Parameterizations. *Educational & Psychological Measurement*, 72 (5).
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods—Advanced quantitative techniques in the social sciences*. Newbury Park, CA: SAGE.
- Cai, L.; Yang, J. S. & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, 16, 221–248. <https://doi:10.1037/a0023350>.
- Chen, T. T. & Fienberg, S. E. (1974). Two-Dimensional Contingency Tables with Both Completely and Partially Cross-Classified Data. *Biometrics*, 30 (4), 629 – 642.
- 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. <https://doi:10.1111/j.1745-3984.2006.00010.x>
- DeMars, C. E. (2013). A Tutorial on interpreting bifactor model scores. *International Journal of Testing*, 13 (4), 354-378. <https://doi:10.1080/15305058.2013.799067>
- Duan, J. C.; Hardle, W. K. & Gentle, J. E. (2012). *Handbook of computational finance*. Springer Heidelberg dordrecht London new york
- Elosua Oleden, P. & Mujika Lizaso, J. (2014). Impact of family language and testing language on reading performance in a bilingual educational context. *Psicothema*, 26 (3), 328-335. <https://doi:10.7334/psicothema2013.344>
- Fukuhara, H. (2009). *A differential item functioning model for testlet-based items using bi-factor multidimensional item response theory*

- model: a bayesian approach, electronic thesis.* treatises and dissertations, florida state university libraries.
- Gibbons, R. D. & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57 (3), 423-436.
- Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications, Inc.
- Holzmüller, H. H.; Singh, J. & Nijssen, E. J. (2002). *Multicetric cross-national research: A typology and illustration*. Retrieved Februray 19, 2015, from <http://www.wiso.tudortmund.de/wiso/m/Medienpool/Arbeitspapier/Arbeitsbericht06.pdf>
- Kim, S. & Kolen, M. J. (2006). Robustness of Format Effects of IRT Linking Methods for Mixed Format Tests. *Applied Measurement in Education*, 19 (4), 357-381.
- Lee, Y.-W. (2004). *Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test*, *Language Testing* (Vol. 21, pp. 74-100). Princeton, NJ: Educational Testing Service.
- 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. <https://doi:10.22237/jmasm/1225512780>.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsboro, NJ: Erlbaum.
- MD Desa, Z. N. D. (2012). *Bi-factor Multidimensional Item Response Theory Modeling for Subscores Estimation, Reliability, and Classification*. Retrieved out 8, 2016, from [https://kuscholarworks.ku.edu/bitstream/handle/1808/10126/MdDesa\\_ku\\_0099D\\_12360\\_DATA\\_1.pdf;sequence=1](https://kuscholarworks.ku.edu/bitstream/handle/1808/10126/MdDesa_ku_0099D_12360_DATA_1.pdf;sequence=1)
- Moore, D. White (2015). *Unidimensional vertical scaling of mixed format tests in the presence of item format effect*, thesis and dissertation, University of Pittsburgh.
- Morton, J. B. & Harper, S. N. (2007). What did Simon say? Revisiting the bilingual advantage. *Developmental Science*, 10, 719-726.
- Mullis, I. V. S.; Martin, M. O.; Gonzalez, E. J., & Kennedy, A. M. (2003). *PIRLS 2001 international report: IEA's study of reading literacy achievement in primary school in 35 countries*. Retrieved on oct 23, 2017. From [https://timss.bc.edu/pirls2001i/pdf/p1\\_IR\\_book.pdf](https://timss.bc.edu/pirls2001i/pdf/p1_IR_book.pdf)
- Mullis, I. V. S.; Martin, M. O.; Kennedy, A. M. & Foy, P. (2007). *PIRLS 2001 international report: IEA's progress in international reading literacy study in primary school in 40 countries*. Retrieved on oct 23, 2017. From [https://timss.bc.edu/PDF/PIRLS2006\\_international\\_report.pdf](https://timss.bc.edu/PDF/PIRLS2006_international_report.pdf)

- Mullis, I. V. S.; Martin, M. O.; Foy, P. & Drucker, K. T. (2012). *PIRLS 2011 international result in reading*, Retrieved on oct 23, 2017. from [https://timssandpirls.bc.edu/pirls2011/downloads/P11\\_IR\\_FullBook.pdf](https://timssandpirls.bc.edu/pirls2011/downloads/P11_IR_FullBook.pdf)
- 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.
- 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.
- Ravand, H. (2015). *Assessing Testlet Effect, Impact Differential Testlet, and Item Functioning Using Cross-Classified Multilevel Measurement Modeling*. SAGE Open 5, 1-9. <https://doi:10.1177/2158244015585607>.
- Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, 28 (3), 237-247.
- Smarter balanced assessment Consortium: 2013-2014 technical report (2015) *Validity, item and test development, pilot test and field test, achievement level setting*. Retrieved February 18, 2015, from [http://www.smarterbalanced.org/wpcontent/uploads/2015/08/201314\\_Technical\\_Report.pdf](http://www.smarterbalanced.org/wpcontent/uploads/2015/08/201314_Technical_Report.pdf)
- Syhabuddin, K. (2013). *Student English Achievement, Attitude and Behaviour in Bilingual and Monolingual Schools in Aceh, Indonesia*. School of Education, Faculty of Education and Arts Edith Cowan University, Perth, Western Australia
- Tao, W. (2008). *Using the score-based testlet method to handle local item dependence*. electronic thesis and dissertation, Boston college.
- Thissen, D.; Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, 26 (3), 247-260.
- Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *The Journal of Educational Psychology*, 16 (7), 433-451.
- van de Vijver, F. J. R., & Leung, K. (1997). *Cross-cultural psychology series, Vol. 1. Methods and data analysis for cross-cultural research*. Thousand Oaks, CA, US: Sage Publications, Inc.
- Wainer, H. & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, 6 (2), 103-118.

- Wang, X.; Bradlow, E. T., & Wainer, H. (2002). A general Bayesian model for testlet: Theory and applications. *Applied Psychological Measurement*, 26 (1), 109-128.
- Wei, L. (2010). BAMFLA: Issues, methods and directions. *International Journal of Bilingualism*, 14 (1), 3-9
- Yao, C. (2008). *Mixed-format test equating: effects of test dimension and common-item sets*. Unpublished Doctoral Dissertation, University of Maryland, College Park.
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30 (3), 187-213.
- Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing*, 27, 119-140. doi:10.1177/0265532209347363.
- Zhang, O.; Shen, L., & Cannady, M.(2010, April). Polytomous IRT or testlet model: An evaluation of scoring models in small testlet size situations. Paper presented at the *Annual Meeting of the 15th International Objective Measurement Workshop*, Boulder, CO, USA.
- Zenisky, A. L., Hambleton, Z. R. K. and Sireci, S.G. (2003). *Effects of local item dependence on the validity of IRT item, Test, and Ability statistics*. Retrieved out 8, 2016, from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.134.5458&rep=rep1&type=pdf>