فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی
سال نهم، شماره ۲۷، پاییز ۱۳۹۸، صص ۷۳-۴۱

Journal of Educational Measurement & Evaluation Studies
Vol. 9, No. 27, Autumn 2019

سنجش برازش فرد در آزمون ریاضی پایه هشتم مطالعه تیمز ۲۰۱۵[۱]

پوریا رضاسلطانی*
ابراهیم خدایی**
جلیل یونسی***
امین موسوی****
علی مقدم‌زاده*****

تاریخ دریافت مقاله: ۹۶/۰۸/۲۲
تاریخ پذیرش مقاله: ۹۷/۰۲/۰۳

چکیده:
روایی نمره‌های حاصل از یک آزمون ممکن است به دلیل رفتارهای پاسخ‌دهی نابهنجار به مخاطره بیافتد. در این پژوهش با استفاده از آماره‌های برازش فرد H^T و l_z^* الگوی پاسخ آزمون ریاضی پایه هشتم مطالعه تیمز دانش‌آموزان کشورهای استرالیا، ایران و جمهوری کره بررسی شده است. پس از تعیین دانش‌آموزان با الگوی پاسخ با برازش نامناسب، تأثیر پاسخ آنها بر برآورد پارامترهای سؤال‌های آزمون قبل و بعد از حذف این دانش‌آموزان مقایسه شده است. تغییرات برآورد پارامترهای برخی از پرسش‌ها قابل ملاحظه بود. همچنین همسانی تعیین الگوی پاسخ دانش‌آموزان با دو آماره برازش فرد H^T و l_z^* صورت گرفت و مشخص شد بین نتایج دو آماره مذکور همسانی وجود دارد. بررسی رابطه الگوی پاسخ با برازش مناسب/ نامناسب دانش‌آموزان و برآورد توانایی آنها نیز انجام گرفت و مشخص شد رابطه معنی‌داری بین برآورد پارامتر توانایی دانش‌آموزان و الگوی پاسخ آنها برای آماره برازش فرد H^T وجود دارد ولیکن برای آماره برازش فرد l_z^* این رابطه معنی‌دار نبود.

واژگان کلیدی: آماره‌های برازش فرد، الگوی پاسخ، رفتارهای پاسخ‌دهی نابهنجار، روایی نمره‌های آزمون، سنجش برازش فرد.

---

[۱]. این مقاله از بخشی از رساله دکتری نویسنده مسئول استخراج شده است.

* دانشجوی دکتری تخصصی رشته سنجش و اندازه‌گیری دانشگاه تهران (نویسنده مسئول: p_rsoltani@ut.ac.ir)
** دانشیار دانشگاه تهران
*** دانشیار دانشگاه علامه طباطبایی
**** استادیار دانشگاه ساسکاچوان کانادا
***** استادیار دانشگاه تهران

مقدمه

استفاده از آزمون‌ها در زمینه‌هایی چون مسائل آموزشی، روان‌شناسی و تصمیم‌سازی، از فعالیت‌های حیاتی است. بر اساس آزمون‌ها تصمیمات خطیری در مورد افراد اتخاذ می‌شود (آلبرس، میجر و تندیرو[1] ۲۰۱۶). وقتی دانش‌آموزان در آزمونی شرکت می‌کنند و به پرسش‌های آزمون پاسخ می‌دهند، پاسخ‌های هر دانش‌آموز به پرسش‌ها به‌اصطلاح الگوی پاسخ دانش‌آموز نامیده می‌شود. الگوهای پاسخ دانش‌آموزان می‌توانند «قابل انتظار»[2] یا «غیر قابل انتظار»[3] باشند. برای توضیح بیشتر الگوهای پاسخ قابل انتظار و غیر قابل انتظار، مثال فرضی ذیل ارائه می‌شود.

فرض کنید یک آزمون پیشرفت تحصیلی شامل ده پرسش است که تمامی ویژگی‌های لازم یک آزمون برای استفاده در پژوهش‌های حوزه تعلیم و تربیت را داشته باشد. به‌طور مثال، دو دانش‌آموز را که نمره یکسان ۵ از ۱۰ کسب کرده‌اند، الگوی پاسخ آنها مورد بررسی قرار می‌گیرد. پارامترهای پرسش‌ها و الگوهای پاسخ‌ها برای مثال فرضی به شرح جدول (۱) است.

جدول (۱) پارامتر پرسش‌ها و الگوی پاسخ‌ها برای مثال فرضی

	سؤال‌ها									
	۱	۲	۳	٤	٥	٦	۷	۸	۹	۱۰
ضریب تشخیص	۱/٦۷	۱/۰۰	۱/۱٤	۱/۳٤	۱/۲۷	۱/۵۰	۱/۸۷	۱/۱۵	۱/۰۰	۱/۸۰
ضریب دشواری	-۲/۰۰	-۱/۵۹	-۰/۸۵	-۰/۱۰	۰/۰۰	۰/۵۰	۱/۲۰	۱/۹۰	۲/۲۰	۲/۵۰
ضریب حدس	۰/۰۱	۰/۲۰	۰/۱۵	۰/۱۵	۰/۱۰	۰/۲۵	۰/۲۰	۰/۱۱	۰/۰۵	۰/۰۱
الگوهای پاسخ										
دانش‌آموز ۱	۱	۱	۱	۱	۱	۰	۰	۰	۰	۰
دانش‌آموز ۲	۰	۰	۰	۰	۰	۱	۱	۱	۱	۱

پارامتر پرسش‌ها بر اساس تئوری پرسش- پاسخ برآورد شده‌اند.

[1]. Albers, Meijer & Tendeiro
[2]. expected
[3]. unexpected

در جدول (۱) ضرایب تشخیص، دشواری و حدس پرسش‌ها و الگوهای پاسخ دو دانش‌آموز ارائه شده است. همچنین پرسش‌ها بر اساس ضریب دشواری‌شان (به‌صورت صعودی) مرتب شده‌اند. مطابق جدول ۱، دانش‌آموز ۱، به پنج پرسش اول آزمون، پاسخ درست و به پنج پرسش آخر پاسخ نادرست داده است؛ به این معنا که دانش‌آموز ۱ در پاسخ درست به پنج پرسش دشوار موفق نبوده است. در مقابل، دانش‌آموز ۲، به پنج پرسش دوم پاسخ درست و به پنج پرسش اول پاسخ نادرست داده است؛ این بدین معناست که دانش‌آموز ۲ در جواب پنج پرسش دشوار موفق بوده ولی در جواب پنج پرسش آسان موفق نبوده است. این در حالی است که هر دو دانش‌آموز ۱ و ۲، نمره یکسان ۵ از ۱۰ را کسب کرده‌اند. الگوی پاسخ دانش‌آموز ۲، «غیر قابل انتظار» است. الگوی پاسخ غیر قابل انتظار در این مثال فرضی با پایش الگوهای پاسخ دانش‌آموزان مشخص شد. آیا در آزمون‌های واقعی این کار عملی است؟ حال چگونه می‌توان دانش‌آموزان با الگوی پاسخ غیر قابل انتظار را شناسایی کرد. این موضوع از مهم‌ترین مسائل حوزه سنجش است که امروزه با عنوان «برازش فرد»[1] شناخته می‌شود.

روش‌های سنجش برازش پاسخ‌های سؤال افراد با مدل اندازه‌گیری آزمون، اغلب به‌عنوان آماره‌های برازش فرد تلقی می‌شوند. آماره‌های برازش فرد میزان تطابق بردار پاسخ فرد با الگوهای پاسخ مورد انتظار فرد در یک مدل خاص را ارزشیابی می‌کنند. در برخی از متون، الگوی پاسخ غیر قابل انتظار، «نابهنجار»[2]، یا «برازش نامناسب»[3] تلقی می‌شوند. برازش نامناسب به تطابق نداشتن الگوهای پاسخ مشاهده شده و مورد انتظار افراد در یک مدل اندازه‌گیری معین اشاره می‌کند. برای سنجش تطابق الگوی پاسخ فرد با یک مدل اندازه‌گیری از یک آماره برازش فرد استفاده می‌شود. در اندازه‌گیری آموزشی و روان‌شناسی، آماره‌های برازش فرد متعددی برای شناسایی افراد با الگوی پاسخ نامناسب عنوان شده است (سویی و لی[4] ۲۰۱۵). به‌طور کلی با آماره‌های برازش فرد، افراد به دو گروه افراد با بردارهای پاسخ با برازش مناسب و افراد با بردارهای پاسخ با برازش نامناسب تقسیم می‌شوند. یکی از مزایای آماره‌های

---

1. person-fit
2. aberrant
3. misfitting
4. Cui & Li

برازش فرد این است که آنها الگوهای پاسخ تک‌تک افرادی را که در آزمون شرکت کرده‌اند، تحلیل می‌کنند. با وجود برچسب‌های متفاوت برای پاسخ‌دهی نابهنجار، به‌طور کلی می‌توان آنها را به دو دسته تقسیم کرد: دسته اول، رفتارهایی که به نمره‌های ساختگی پایین[1]منجر می‌شوند و دسته ، رفتارهایی که به نمره‌های ساختگی بالا[2]منجر می‌شوند. وقتی که در آزمونی هر دو نوع این رفتارها رخ دهد به آن نمره‌های ساختگی آمیخته[3]گویند (راپ[4]، ۲۰۱۳).

الگوی پاسخ غیر قابل انتظار در برآورد توانایی دانش‌آموزان، پارامتر پرسش‌ها و سایر ویژگی‌های روان‌سنجی آزمون بی‌تأثیر نخواهد بود. کونیجن، ایمونز و سیجتسما[5] (۲۰۱٤)، میجر[6](۱۹۹۷) و اشمیت، کورتینا و ویتنی[7](۱۹۹۳) بیان کردند روایی و پایایی یک آزمون ممکن است به دلیل وجود پاسخ‌های با برازش نامناسب برخی از افراد (نه همه افراد) به مخاطره بیفتد.

مطالعه تیمز یکی از ملاک‌های ارزشیابی نظام‌مند درون‌دادها، فرایندها و برون‌دادهای نظام آموزشی ایران در سطح ملی، منطقه‌ای و جهانی است. برای انجام مطالعه تیمز هزینه‌های زیادی صرف می‌شود، با استفاده از تکنیک تکمیلی برازش فرد در تحلیل نتایج این مطالعه، نتایج قابل اعتمادتری برای پارامترهای پرسش‌ها و توانایی دانش‌آموزان به دست خواهد آمد. اصلاح در برآورد توانایی دانش‌آموزان با استفاده از تکنیک برازش فرد، می‌تواند شاخص‌های متعدد مطالعه تیمز برای کشور ایران و سایر کشورها را تحت تأثیر قرار دهد و ممکن است رده‌بندی آنها را نیز جابجا کند. هدف کلی از اجرای این پژوهش، بررسی تأثیر الگوهای پاسخ با برازش نامناسب بر برآورد توانایی دانش‌آموزان و پارامترهای پرسش‌های آزمون است. هدف‌های اختصاصی در این پژوهش، مقایسه برآورد پارامتر پرسش‌های آزمون، پیش و پس از حذف دانش‌آموزان با الگوی پاسخ با برازش نامناسب؛ مقایسه بین مقدار آماره‌های برازش

---

1. spuriously low scores
2. spuriously high scores
3. spuriously mixed scores
4. Rupp
5. Conijn, Emons & Sijtsma
6. Meijer
7. Schmitt, Cortina & Whitney

فرد دانش‌آموزان؛ و بررسی رابطه الگوی پاسخ دانش‌آموزان با توانایی آنهاست. پرسش‌های پژوهش عبارت‌اند از:

-پارامترهای پرسش‌ها پیش و پس از حذف دانش‌آموزان با الگوی پاسخ نامناسب چقدر است؟

-آیا تعیین الگوی پاسخ دانش‌آموزان با استفاده از آماره‌های برازش فرد H^T و l_z^* یکسان است؟

-آیا رابطه‌ای بین برآورد توانایی دانش‌آموزان با الگوی پاسخ آنها وجود دارد؟

مبانی نظری و پیشینه پژوهش

آماره‌های برازش فرد را می‌توان با دو رویکرد اصلی طبقه‌بندی کرد: آماره‌های برازش فرد گروه- مبنا[1] و آماره‌های برازش فرد IRT- مبنا. در رویکرد اول (گروه- مبنا)، آماره‌های برازش فرد بدون در نظر گرفتن مدل اندازه‌گیری خاصی و از مقایسه الگوی پاسخ مشاهده شده یک فرد بر اساس مشخصه‌های نمره کل آزمون از روی کل نمونه محاسبه می‌شوند (کاراباتسوس[2] ۲۰۰۳). یک آماره برازش فرد گروه- مبنا، یک الگوی پاسخ مشاهده شده را با عنوان برازش نامناسب طبقه‌بندی می‌کند؛ اگر به پرسش‌های ساده، پاسخ نادرست و به پرسش‌های دشوار، پاسخ درست داده شوند (میجر و سیجتسما[3] ۲۰۰۱). اگر نمره فردی در یک آزمون r باشد، انتظار می‌رود که فرد r تا از ساده‌ترین پرسش‌ها را به‌طور درست پاسخ دهد. یک بردار پاسخ به‌عنوان برازش نامناسب تلقی می‌شود وقتی‌که پرسش‌های به‌طور نسبی با نسبت پاسخ صحیح پایین، درست پاسخ داده شوند ولیکن پرسش‌های به‌طور نسبی با نسبت پاسخ صحیح بالا، نادرست پاسخ داده شوند. نمونه‌هایی از آماره‌های برازش فرد گروه- مبنا عبارت‌اند از شاخص احتیاط تعدیل شده C هارنیش و لین[4] (۱۹۸۱)، شاخص U3

---

1. group-based
2. Karabatsos
3. Sijtsma
4. Harnisch & Linn

ون‌درفلایر[1] (۱۹۸۲)، شاخص انطباق هنجار NCI تاتسوکا و تاتسوکا[2] (۱۹۸۳)، و ضریب H^T سیجتسما[3](۱۹۸٦).

در رویکرد دوم (IRT- مبنا)، آماره‌های برازش فرد، برازش الگوی پاسخ را با یک مدل IRT مشخص از قبیل مدل سه‌پارامتری لجستیک، می‌سنجند. آماره‌های برازش فرد مبتنی بر مدل، با استفاده از برآورد پارامترهای پرسش و توانایی افراد، محاسبه شده و افراد به دو گروه با برازش مناسب و برازش نامناسب طبقه‌بندی می‌شوند. آماره‌های برازش فرد IRT- مبنا به‌طور ویژه برای ارزشیابی برازش نامناسب از روی الگوی پاسخ مشاهده شده با مدل IRT به‌وسیله محاسبه احتمال‌های پاسخ مرتبط با پارامترهای پرسش و پارامتر توانایی افراد طراحی شدند. اگر مطابق مدل IRT، احتمال پاسخ درست فرد بالا باشد، فرضیه این است که فرد باید به پرسش، پاسخ درست بدهد و برعکس. یک برازش نامناسب وقتی رخ می‌دهد که فرضیه فوق توسط داده‌ها تأیید نشود. نمونه‌هایی از آماره‌های برازش فرد IRT- مبنا عبارت‌اند از آماره U رایت و استون[4](۱۹۷۹)، آماره W رایت و مسترز[5](۱۹۸۲)، آماره‌های UW و UB اسمیت[6] (۱۹۸۵)، آماره l_z دراسگو، لوین، و ویلیامز[7](۱۹۸۵)، و آماره l_z^* اسنیجرز[8](۲۰۰۱). با توجه به مقاله کاراباتسوس (۲۰۰۳)، آماره برازش فرد H^T، بهترین عملکرد در بین آماره‌های برازش فرد دارد و آماره برازش فرد l_z، نیز از بین آماره‌های برازش فرد IRT- مبنا، عملکرد بهتری دارد. در ارتباط با آماره برازش فرد l_z، به‌دلیل اینکه توزیع آن وقتی پارامتر توانایی افراد برآورد می‌شود، توزیع نرمال نیست، اسنیجرز (۲۰۰۱)، آماره برازش فرد l_z^* که نسخه تعدیل شده l_z است را ارائه کرد. بنابراین در این پژوهش از دو آماره برازش فرد H^T و l_z^* استفاده و در ادامه تشریح می‌شود.

---

1. van der Flier
2. Tatsuoka & Tatsuoka
3. Sijtsma
4. Wright & Stone
5. Wright & Masters
6. Smith
7. Drasgow, Levine & Williams
8. Snijders

آماره برازش فرد H^T

در اغلب آماره‌های برازش فرد گروه–مبنا (ناپارامتریک)، الگوی پاسخ مشاهده شده با الگوی پاسخ مورد انتظار تحت مدل گاتمن (گاتمن[1] ۱۹۴۴، به نقل از موسوی[2] ۲۰۱۵) مقایسه می‌شوند. مدل گاتمن عبارت است از:

$$\theta_i < \delta_j \quad \leftrightarrow \quad P_j(\theta_i) = 0,$$
$$\theta_i \geq \delta_j \quad \leftrightarrow \quad P_j(\theta_i) = 1$$

که در آن $P_j(\theta_i)$ احتمال پاسخ درست دانش‌آموز i–ام با توانایی θ_i به سؤال j–ام با ضریب دشواری δ_j است. بر اساس این مدل، دانش‌آموزی با سطح توانایی θ می‌تواند به r تا از پرسش‌های با ضریب دشواری کمتر یا مساوی با θ پاسخ درست بدهد و به J-r تا از سؤال‌های با ضریب دشواری بیشتر از θ پاسخ نادرست بدهد، و در آن J تعداد پرسش‌هاست. چنین الگوی پاسخی به «الگوی گاتمن» یا «الگوی سازگار» مشهور است. اگر دانش‌آموزی به J-r پرسشی که ضریب دشواری بیشتر از θ دارد، پاسخ درست بدهد و در پاسخ دادن به r پرسش با ضریب دشواری کمتر یا مساوی θ ناموفق باشد، الگوی پاسخ او با عنوان «خطا» یا «وارونه» یا «الگوی گاتمن معکوس» در نظر گرفته می‌شود (میجر و سیجتسما، ۲۰۰۱).

آماره برازش فرد H^T با فرمول ذیل محاسبه می‌شود (موسوی، ۲۰۱۵).

$$H^T = \frac{\sum_{i \neq j} \sigma_{ij}}{\sum_{i \neq j} \sigma_{ij}^{max}}$$

که در آن σ_{ij} کوواریانس بین بردارهای پاسخ دانش‌آموزان i و j است و σ_{ij}^{max} ماکسیمم ممکن کوواریانس بین بردارهای پاسخ دانش‌آموزان i و j است. با توجه به فرمول بالا، آماره برازش فرد H^T همبستگی بین بردار پاسخ مشاهده‌شده دانش‌آموز i با سایر دانش‌آموزان را می‌سنجد. مقدار آماره برازش فرد H^T بین ۱ و ۱– می‌تواند باشد. به عبارت دیگر، آماره برازش فرد H^T بردار پاسخ یک دانش‌آموز را با بردار پاسخ سایر دانش‌آموزان نمونه مقایسه می‌کند. اگر بردار پاسخ مشاهده‌شده یک دانش‌آموز، با الگوهای پاسخ سایر دانش‌آموزان سازگار باشد، آنگاه صورت کسر آماره

---

[1]. Guttman
[2]. Mousavi

برازش فرد H^T و درنتیجه مقدار آماره برازش فرد H^T مثبت می‌شود. به همین ترتیب، اگر بردار پاسخ مشاهده‌شده یک دانش‌آموز، با الگوهای پاسخ سایر دانش‌آموزان سازگار نباشد، آنگاه صورت کسر آماره برازش فرد H^T و درنتیجه مقدار آماره برازش فرد H^T منفی می‌شود. بردار پاسخ تصادفی که با سایر الگوهای پاسخ ناهمبسته باشد، به مقدار صفر برای آماره برازش فرد H^T منجر می‌شود (میجر و سیجتسما، ۲۰۰۱). سیجتسما و میجر (۱۹۹۲) مقدار برش ۰/۳ برای آماره برازش فرد H^T را پیشنهاد دادند. الگوهای پاسخ با مقدار H^T کمتر از ۰/۳ به‌عنوان برازش نامناسب تلقی می‌شوند.

آماره برازش فرد l_z^*

آماره برازش فرد l_z^* (اسنیجرز، ۲۰۰۱) نسخه تعدیل‌شده آماره برازش فرد l_z (دراسگو و همکاران، ۱۹۸۵) است که به‌منظور مرتفع کردن مشکل توزیع آماره برازش فرد l_z هنگامی که پارامتر توانایی افراد برآورد می‌شود، معرفی شد. فرمول محاسبه آماره برازش فرد l_z^* به شرح ذیل برای تمامی دانش‌آموزان محاسبه می‌شود، جزئیات فرمول محاسبه آماره‌ی برازش فرد l_z^* در مجیس[1] رایچ[2] و بلند[3](۲۰۱۲) آمده است.

$$l_z^* = \frac{l_0(\hat{\theta}) - E[l_0(\hat{\theta})] + c_n(\hat{\theta})r_0(\hat{\theta})}{\tilde{V}[l_0(\hat{\theta})]^{1/2}},$$

$$\tilde{V}[l_0(\theta)] = \sum_{i=1}^{n} \tilde{w}_i(\theta)^2 P_i(\theta) Q_i(\theta),$$

به‌دلیل اینکه مقادیر بزرگ منفی آماره برازش فرد l_z^* نشان‌دهنده برازش نامناسب بالقوه است، نقطه برش برای تعیین برازش نامناسب از دنباله سمت چپ توزیع نرمال استاندارد با انتخاب یک سطح خطاپذیری (به‌طور مثال، $\alpha = 0.05$) حاصل می‌شود.

---

1. Magis
2. Raich
3. Bland

تأثیر برازش نامناسب فرد بر مشخصه‌های پرسش و آزمون

همان‌طور که میجر و سیجتسما (۲۰۰۱) گزارش کرده‌اند، در مطالعات اندکی اثر الگوهای پاسخ با برازش نامناسب روی مشخصه‌های آزمون و پارامترهای پرسش بررسی شده است. لوین و دراسگو (۱۹۸۲) از مدل لجستیک سه‌پارامتری برای بررسی تأثیر استفاده از برآورد پارامترهای پرسش به‌جای مقادیر واقعی پارامترهای پرسش روی آماره برازش فرد l_0، و تأثیر الگوهای پاسخ با برازش نامناسب روی آماره برازش فرد l_0 و برآورد پارامترهای پرسش بهره برده‌اند. پارامترهای پرسش برآورد شده از مدرج‌سازی[1] قبلی بخش کلامی آزمون استعداد تحصیلی به‌منظور تولید داده‌های شبیه‌سازی استفاده شده است. لوین و دراسگو (۱۹۸۲) نتیجه گرفتند که وجود الگوهای پاسخ با برازش نامناسب تأثیری روی آماره برازش فرد l_0 و برآورد پارامتر پرسش ندارد. آنها استدلال کردند پاسخ‌های با برازش نامناسب متفاوت به داشتن الگوهای پاسخ نادرست متفاوت تمایل دارند و درنتیجه تعداد زیادی از الگوهای پاسخ با برازش نامناسب اثرات متضادی بر برآورد پارامترهای سؤال دارند.

فیلیپس[2] (۱۹۸۶) تأثیر بردارهای پاسخ با برازش نامناسب بر برازش مدل راش، برآورد پارامترهای سؤال، و همترازسازی[3] با صدک‌های یکسان را بررسی کرد. وی متوجه شد که حذف الگوهای پاسخ با برازش نامناسب می‌تواند: ۱- برازش مدل با داده‌ها را بهبود بخشد؛ ۲- تأثیر اندکی بر برآورد پارامتر دشواری پرسش‌ها داشته باشد؛ ۳ اساساً تأثیری بر نتایج همترازسازی نداشته باشد. در مطالعه دیگری، رادنر، بریسی و اسکاگ[4] (۱۹۹۶) داده‌های سنجش ملی پیشرفت تحصیلی (NAEP) سال ۱۹۹۰ را تحلیل کردند و تقریباً هیچ بردار پاسخ با الگوی نامناسبی پیدا نکردند. درنتیجه تفاوت معنی‌داری در میانگین آزمون پیش و پس از حذف الگوهای پاسخ با برازش نامناسب، دیده نشد.

سوتاریدنا، چوی و میجر[5] (۲۰۰۵) اثر الگوهای پاسخ با برازش نامناسب بر مدرج‌سازی پرسش و طبقه‌بندی عملکرد را مطالعه کردند. آنها از آماره‌های آزمون l_z^*

---

1. calibration
2. Phillips
3. equating
4. Rudner, Braccy & Skagg
5. Sotaridona, Choi & Meijer

و U3 به‌عنوان آماره‌های برازش فرد و یک نمونه تصادفی ۱۰هزارتایی دانش‌آموزان از یک برنامه سنجش سراسری ریاضی و علوم استفاده کردند. هر آزمون شامل ۵۵ پرسش چندگزینه‌ای بود. دو نوع پاسخ‌دهی با برازش نامناسب (کپی کردن و حدس زدن) با دستکاری داده‌ها از دانش آموزان منتخب، شبیه‌سازی شده بود. درنتیجه، دو مجموعه داده‌ها شبیه‌سازی شدند- مجموعه داده‌های اصلی و مجموعه داده‌های با پاسخ‌های با برازش نامناسب شبیه‌سازی شده. مجموعه داده‌ها به‌طور مستقل با استفاده از مدل لجستیک سه‌پارامتری مدرّج شدند و برآورد پارامترهای پرسش با روش استوکینگ و لرد[1](۱۹۸۳) همترازسازی شدند. چهار ملاک برای مقایسه تفاوت در برآورد پارامترهای همترازسازی شده، تفاوت در خطاهای استاندارد برآورد پارامترها، تفاوت در منحنی‌های مشخصه آزمون، و تفاوت در منحنی‌های آگاهی آزمون، انتخاب شدند. نتایج نشان داد که با وجود برازش نامناسب فرد، پارامترها به‌طور قابل توجهی بیش‌برآورد شدند و خطای استاندارد برآورد برای داده‌های با الگوهای پاسخ با برازش نامناسب بالاتر بودند. منحنی‌های مشخصه و آگاهی آزمون به‌طور معنی‌داری متفاوت نبودند. علاوه بر مجموعه داده‌های اصلی، دو مجموعه داده‌های دیگر تولید شدند؛ یک مجموعه داده‌ها شامل پاسخ‌های با برازش مناسب با استفاده از آماره برازش فرد l_z^* و یک مجموعه داده‌ها شامل پاسخ‌های با برازش مناسب با استفاده از آماره برازش فرد U3. هر مجموعه داده‌ای به‌طور مستقل، مدرّج و همترازسازی شدند؛ نمره‌های مقیاس استاندارد شده به سه سطح، مهارت کمتر، ماهر و پیشرفته تبدیل شدند. نتایج نشان‌دهنده این بود که تفاوت‌ها ناچیز هستند و در مجموعه داده‌های منتخب با استفاده از آماره برازش فرد U3، تفاوت‌ها عموماً کوچک‌تر بودند. آنها نتیجه گرفتند که شمول الگوهای پاسخ با برازش نامناسب، دقت برآورد پارامترها را کاهش داده است، ولیکن در سطح آزمون اثرات حداقل بوده‌اند.

هندراوان، گلاس و میجر (۲۰۰۵) نیز تأثیر الگوهای پاسخ با برازش نامناسب بر تصمیم‌های طبقه‌بندی را بررسی کردند. آنها از سه روش برآورد توانایی-MLĔ[2]، EAP̆[3]، و MCMC̓[4]- با مدل اجایو نرمال سه‌پارامتری استفاده کردند و از

---

[1]. Stocking & Lord
[2]. maximum likelihood estimate
[3]. expected a posterior
[4]. Markov Chain Monte Carlo

برآوردگرهای ماکسیمم درست‌نمایی حاشیه‌ای (MML[1]) و روش بیزی برای برآورد پارامترهای پرسش بهره بردند. از پنج آماره برازش فرد W (رایت و استون، ۱۹۷۹)، UB (اسمیت، ۱۹۸٦)، ζ_1 و ζ_2 (تاتسوکا، ۱۹۸٤) و l_z استفاده شد. در شبیه‌سازی دو آزمون ۳۰ و ٦۰ سؤالی، دو نوع پاسخ با برازش نامناسب (حدس زدن و افشای پرسش)، سه مقدار ضریب تشخیص پرسش (۰/۵، ۱ و ۱/۵)، دو اندازه نمونه (٤۰۰ و ۱۰۰۰)، و سه مقدار برش برای تعیین تسلط/ عدم تسلط در آزمون (۱–، ۰ و ۱) بررسی شدند. نتایج نشان داد که وجود الگوهای پاسخ با برازش نامناسب به برآوردهای اریب پارامترهای پرسش، تصمیم‌های نادرست طبقه‌بندی تسلط، مخصوصاً برای رفتار حدس زدن که میانگین توزیع توانایی برآورد شده را کمتر نشان می‌دهد، منتج می‌شود. اینها به‌طور مصنوعی به‌دقت طبقه‌بندی بالاتر برای دانش‌آموزانی با توانایی پایین به‌دلیل بیش–برآورد پارامتر حدس و کم–برآورد پارامتر دشواری سؤال، منجر شدند. در مورد افشای پرسش که دانش‌آموز پیش–آگاهی از پرسش‌ها دارد، نتایج با رفتار مورد انتظار متناقض بود. دقت طبقه‌بندی برای دانش‌آموزان با توانایی بالا بیشتر بود، به‌دلیل اینکه افشای پرسش به بیش–برآورد توانایی منجر شده بود. تمامی آماره‌های برازش فرد به‌خوبی عمل کرده و موجب طبقه‌بندی بهتر شده است. به‌طور کلی برای تمامی روش‌های برآورد مذکور، نتایج یکسان بودند و MCMC از نظر دقت طبقه‌بندی برای الگوهای پاسخ با برازش مناسب و الگوهای پاسخ با برازش نامناسب، بدترین حالت بود. به‌عنوان یک نتیجه‌گیری کلی، هندراوان و همکاران (۲۰۰۵) استدلال کردند که آماره‌های برازش فرد در پیدا کردن زیرنمونه‌های با برازش مناسب، مفید و برای استفاده در آزمون تسلط، مناسب هستند.

موسوی (۲۰۱۵)، اثرات شمول و عدم شمول الگوهای پاسخ با برازش نامناسب بر برآوردهای پارامتر پرسش با استفاده از داده‌های شبیه‌سازی را بررسی کرد. چهار عامل طول آزمون (۲۰، ٤۰، و ٦۰ پرسش)، روش برآورد پارامتر پرسش (برآورد ماکزیمم درست‌نمایی و روش بیزی)، درصدی از دانش‌آموزان که با برازش نامناسب به پرسش‌ها پاسخ می‌دهند (۱۰٪، ۲۰٪، و ۳۰٪)، و درصدی از پرسش‌های تأثیرپذیر از پاسخ‌های با برازش نامناسب (۲۵٪ و ۵۰٪) در نظر گرفته شد. دو آماره برازش فرد l_z و H^T برای حذف الگوهای پاسخ با برازش نامناسب از مجموعه داده‌ها استفاده شده

[1]. marginal maximum likelihood

است. از مدل لجستیک دوپارامتری نیز برای تحلیل داده‌ها استفاده شده است. متغیرهای وابسته عبارت بودند از اریبی در برآورد پارامترهای دشواری و تشخیص پرسش‌ها، خطای استاندارد برآورد پارامترها، و درستی طبقه‌بندی عملکرد دانش‌آموزان. نتایج نشان داد که ۱- تفاوتی بین دو روش برآورد پرسش وجود نداشت؛ ۲- پارامتر دشواری پرسش نسبت به پارامتر تشخیص پرسش کمتر تحت تأثیر الگوهای پاسخ با برازش نامناسب قرار داشت؛ ۳- پارامترهای پرسش با مقادیر واقعی پارامترهای دشواری و تشخیص بالا، تحت تأثیر الگوهای پاسخ با برازش نامناسب بودند؛ ٤- افزایش در درصد الگوهای پاسخ با برازش نامناسب به اریبی بیشتر در برآورد هر دو پارامتر دشواری و تشخیص پرسش منجر می‌شود؛ ۵- خطاهای استاندارد برآورد پارامترهای دشواری و تشخیص پرسش‌ها در تمامی موارد کوچک بودند؛ ٦- درستی طبقه‌بندی برای دانش‌آموزان با عملکرد پایین برای مجموعه داده‌های با برازش مناسب، بیشتر بود و برای مجموعه داده‌های با الگوهای پاسخ با برازش نامناسب و مجموعه داده‌های با الگوی پاسخ با برازش نامناسب حذف شده توسط دو آماره برازش فرد l_z و H^T کمتر یا ثابت گزارش شده است؛ ۷- درستی طبقه‌بندی برای دانش‌آموزان با عملکرد بالا نسبت به دانش‌آموزان با عملکرد پایین برای مجموعه داده‌های با برازش مناسب پایین‌تر بود و برای سایر سه مجموعه داده‌ها کمتر یا ثابت گزارش شده است.

به‌طور کلی، می‌توان گفت که آماره‌های برازش فرد به‌منظور شاخصی برای شناسایی الگوی پاسخ دانش‌آموزان به‌کار برده می‌شود و از بین آماره‌های برازش فرد موجود، آماره‌های برازش H^T و l_z^* عملکرد بهتری دارند. همچنین الگوی پاسخ نامناسب دانش‌آموزان می‌تواند به تغییر در برآورد پارامترهای پرسش‌ها و برآورد توانایی دانش‌آموزان منجر شود.

روش پژوهش

به استناد طبقه‌بندی کلی پژوهش‌ها به دو گروه آزمایشی و غیرآزمایشی (توصیفی)، این پژوهش از تحقیقات غیرآزمایشی (توصیفی) محسوب می‌شود. همچنین از منظر تقسیم‌بندی تحلیل‌ها به دو نوع اولیه و ثانویه، تحلیل‌های این پژوهش از نوع تحلیل ثانویه هستند.

جامعه آماری این پژوهش، شامل تمامی دانش‌آموزان ثبت نام کرده در پایه هشتم کشورهای ایران، جمهوری کره[۱] و استرالیا[۲] در سال ۲۰۱۵ میلادی (سال تحصیلی ۱۳۹۳-۱۳۹٤) است. متوسط سن دانش‌آموزان در زمان آزمون حداقل ۱۳,۵ سال بود. برخی از دانش‌آموزان که در مدرسه‌های خاصی مشغول به تحصیل بودند، جزء جامعه آماری نیستند؛ دانش‌آموزان مدرسه‌هایی که از لحاظ موقعیت جغرافیایی، دسترسی به آنها آسان نبود؛ دانش‌آموزان مدرسه‌هایی که تعداد دانش‌آموزان پایه هشتم آن مدرسه‌ها خیلی کم (چهار یا کمتر) بود؛ دانش‌آموزان مدرسه‌هایی که از لحاظ ساختار، برنامه درسی و ... از سیستم مدرسه‌های عمومی کاملاً متفاوت بودند و دانش‌آموزان با ناتوانی عملکردی، معلولیت ذهنی و زبان غیربومی جزء جامعه آماری محسوب نمی‌شوند (مارتین، مولیس و هوپر[۳] ۲۰۱٦).

در مطالعه تیمز به‌دلیل ماهیت ذاتی سلسله‌مراتبی دانش‌آموزان در مدرسه، از روش نمونه‌گیری خوشه‌ای دومرحله‌ای استفاده می‌شود. در مرحله اول نمونه‌گیری، مدرسه‌ها با احتمال متناسب با اندازه (PPS[۴]) از لیست که شامل دانش‌آموزان واجد شرایط هستند، انتخاب می‌شوند. در مرحله دوم نمونه‌گیری، یک یا چند کلاس هشتم از هر یک از مدرسه‌های انتخاب شده در مرحله اول نمونه‌گیری انتخاب می‌شود. نمونه‌گیری از کلاس‌ها به‌وسیله هماهنگ‌کننده‌های ملی پژوهش (NRC[۵]) با استفاده از نرم‌افزار نمونه‌گیری درون مدرسه (WinW3S) تولید شده توسط IEA DPC[٦] و مرکز آمار کانادا انجام می‌شود (مارتین و همکاران، ۲۰۱٦).

از آنجایی که مطالعه تیمز اساساً یک مطالعه پیشرفت تحصیلی دانش‌آموزان است، دقت برآوردهای پیشرفت تحصیلی دانش‌آموزان از اهمیت بالایی برخوردار است. برای تأمین کردن استانداردهای مطالعه تیمز در مورد دقت نمونه‌گیری، تعداد نمونه‌های هر کشور باید طوری باشد که خطای استاندارد[۷] میانگین نمره پیشرفت

---

۱. به‌عنوان یک کشور آسیایی با رتبه برتر در مطالعه تیمز

۲. به‌عنوان یک کشور با رتبه متوسط در مطالعه تیمز

3. Martin, Mullis & Hooper
4. Probability Proportional to their Size
5. National Research Coordinator
6. IEA Data Processing and research Center
7. standard error

تحصیلی دانش‌آموزان آن کشور بزرگ‌تر از ۰/۰۳۵ انحراف استاندارد[1] نمره پیشرفت تحصیلی دانش‌آموزان آن کشور نباشد. برای اغلب کشورهای شرکت‌کننده در مطالعه تیمز ۲۰۱۵، انتخاب حداقل ۱۵۰ مدرسه و ٤۰۰۰ دانش‌آموز پایه هشتم، استانداردهای لازم مطالعه تیمز را محقق می‌سازد. در مطالعه تیمز ۲۰۱۵، تعداد مدرسه‌های انتخاب شده برای کشورهای ایران، استرالیا و جمهوری کره به ترتیب ۲۵۰، ۲۸۵ و ۱۵۰ مدرسه است. اندازه نمونه در پایه هشتم برای کشورهای ایران، استرالیا و جمهوری کره به ترتیب ٦۱۳۰، ۱۰۳۳۸ و ۵۳۰۹ دانش‌آموز است (مارتین و همکاران، ۲۰۱٦).

ابزارهای گردآوری داده‌های مطالعه تیمز ۲۰۱۵، از دو بخش آزمون پیشرفت تحصیلی و پرسشنامه‌های زمینه‌ای، تشکیل می‌شود. آزمون پیشرفت تحصیلی ریاضی پایه هشتم، شامل ۲۱۲ پرسش برای اندازه‌گیری دانش و مهارت‌های ریاضی دانش‌آموزان پایه هشتم و همچنین سه پرسشنامه زمینه‌ای شامل پرسشنامه دانش‌آموز، پرسشنامه دبیر ریاضی و پرسشنامه مدیر مدرسه است (کبیری، کریمی و بخشعلی‌زاده، ۱۳۹۵). روایی محتوایی تمامی پرسشنامه‌ها به‌دقت بررسی شده است. برای بررسی روایی ملاکی مقیاس‌های مختلف پرسشنامه‌های زمینه‌ای از ضریب همبستگی بین آن مقیاس‌ها با نمره آزمون پیشرفت تحصیلی ریاضی استفاده شده است؛ به‌طور مثال ضریب همبستگی پیرسون بین مقیاس اعتماد به نفس دانشجویان در ریاضی و نمره آزمون پیشرفت تحصیلی ریاضی برای کشورهای ایران، استرالیا و جمهوری کره به ترتیب ۰/٤۲، ۰/۵۱ و ۰/۵۲ است (مارتین و همکاران، ۲۰۱٦، ۱۵,۲۱۲). برای بررسی پایایی نیز ضریب پایایی آلفای کرونباخ آزمون پیشرفت تحصیلی ریاضی پایه هشتم مطالعه تیمز ۲۰۱۵ برای کشورهای ایران، استرالیا و جمهوری کره به ترتیب عبارت از ۰/۸۷، ۰/۸۹ و ۰/۹۱ بود (بخش‌های ۱۱ و ۱٦ مارتین و همکاران، ۲۰۱٦). ضریب پایایی آلفای کرونباخ مقیاس‌های مختلف پرسشنامه‌های زمینه‌ای در سطح قابل قبولی قرار دارند، تقریباً همه بالای ۰/۷ و بسیاری از آنها بالای ۰/۸ است (مارتین و همکاران، ۲۰۱٦، ۱۵,۱۰).

شایان ذکر است در مطالعه تیمز برای پوشش تمام و کمال برنامه درسی ریاضی از روش نمونه‌گیری ماتریسی پرسش‌ها استفاده می‌شود. به عبارتی، هر یک از

---

[1]. standard deviation

دانش‌آموزان به تمامی پرسش‌های خزانه پرسش‌ها[1] پاسخ نمی‌دهد، بلکه زیرمجموعه‌ای از سؤال‌ها را پاسخ می‌دهد (اولسون، مارتین و مولیس[2] ۲۰۰۸). یکی از رویکردها برای پوشش تمام و کمال برنامه درسی و در عین حال به حداقل رساندن زمان آزمون برای دانش‌آموزان، استفاده از روش نمونه‌گیری ماتریسی پرسش‌ها است. روش نمونه‌گیری ماتریسی پرسش‌ها مشتمل بر توسعه مجموعه کاملی از پرسش‌های در نظر گرفته برای پوشش برنامه درسی است؛ در این روش پرسش‌ها به زیرمجموعه‌هایی تقسیم می‌شوند و هر یک از دانش‌آموزان صرفاً به یکی از زیرمجموعه‌های پرسش‌ها پاسخ می‌دهند. نمونه‌گیری ماتریسی با محدود کردن تعداد پرسش‌ها که هر دانش‌آموز پاسخ می‌دهد، زمان مورد نیاز آزمون را نیز محدود می‌کند؛ این در حالی است که پوشش تمام و کمال محتوای آزمون برای تمامی دانش‌آموزان تعبیه شده است (چایلدز و ژاسیو[3] ۲۰۰۳).

روش تحلیل داده‌ها

برای تجزیه‌وتحلیل داده‌ها، ابتدا با استفاده از نرم‌افزار تحلیل پایگاه داده‌های بین‌المللی IEA IDB[4] (IEA, 2017) داده‌های اولیه مورد نیاز این پژوهش از پایگاه داده‌های مطالعه تیمز ۲۰۱۵ در قالب فایل SPSS استخراج شده است. سپس برای بررسی پیش‌فرض‌های مدل از نرم‌افزار NOHARM استفاده شد. پس از بررسی پیش‌فرض‌ها و انتخاب مدل، پارامترهای پرسش‌ها و توانایی دانش‌آموزان با استفاده از نرم‌افزار BILOG-MG برآورد می‌شوند. برای محاسبه مقادیر آماره‌های برازش فرد از بسته نرم‌افزاری PerFit موجود در نرم‌افزار R استفاده می‌شود (موسوی، تندیرو و یونسی[5] ۲۰۱۶). برای تجزیه‌وتحلیل داده‌ها و پاسخ به پرسش‌های پژوهش با استفاده از نرم‌افزار SPSS از روش‌های آماری ارائه آماره‌های توصیفی و آزمون مک‌نمار، آزمون خی‌دو، و آزمون دقیق فیشر استفاده شده است.

---

1. Item pool
2. Olson, Martin & Mullis
3. Childs & Jaciw
4. International Database Analyzer
5. Mousavi, Tendeiro & Younesi

یافته‌های پژوهش

در این قسمت به ترتیب هر یک از پرسش‌های پژوهش، نتایج آن ارائه می‌شود.

پرسش اول: پارامترهای پرسش‌ها پیش و پس از حذف دانش‌آموزان با الگوی پاسخ نامناسب چقدر است؟

برای به دست آوردن پارامترهای پرسش‌های آزمون، ابتدا مفروضه اصلی تئوری پرسش- پاسخ، تک‌بعدی بودن[1] بررسی می‌شود. تک‌بعدی بودن، بدین معناست که تمام پرسش‌های آزمون، تنها یک صفت یا ویژگی مکنون را اندازه‌گیری می‌کنند.

روش‌های متعددی چون تحلیل عاملی خطی[2] تحلیل عاملی غیرخطی[3] و تحلیل عاملی با اطلاعات کامل[4] برای بررسی مفروضه تک‌بعدی بودن مجموعه پرسش‌های آزمون ارائه شده است (مینایی و فلسفی‌نژاد، ۱۳۸۹). نتایج پژوهش‌های متعدد نشان داده است که تحلیل عاملی خطی برای سنجش ابعاد داده‌های پرسش‌های چندگزینه‌ای مناسب نیست. از بین تحلیل عاملی غیرخطی و تحلیل عاملی با اطلاعات کامل نتایج برخی از پژوهش‌ها (دی‌چمپلین و گساروِلی[5] ۱۹۹۸؛ فینچ و هابینگ[6] ۲۰۰۵؛ نول و برگر[7] ۱۹۹۱) نشان داده‌اند که در چنین مطالعاتی تحلیل عاملی غیرخطی بهتر از تحلیل عاملی با اطلاعات کامل عمل می‌کند. از این‌رو در پژوهش حاضر از روش تحلیل عاملی غیرخطی با استفاده از نرم‌افزار NOHARM برای بررسی مفروضه تک‌بعدی بودن استفاده شد. در جدول (۲) شاخص‌های محاسبه شده توسط NOHARM برای بررسی تک‌بعدی بودن داده‌ها ارائه می‌شود.

---

1. unidimensionality
2. linear factor analysis
3. non-linear factor analysis
4. full information factor analysis
5. De Champlain & Gessaroli
6. Finch & Habing
7. Knol & Berger

جدول (۲) شاخص‌های بررسی تک‌بعدی بودن داده‌ها

تعداد ابعاد	تعداد سؤال‌ها	تعداد آزمودنی‌ها	مجموع مجذورات مانده‌ها	ریشه دوم میانگین مجذورات مانده‌ها	شاخص برازش تاناکا
تک‌بعدی	۲۹	۱۵۷۹	۰/۰۲۵۷۷۷۳	۰/۰۰۷۹۶۸۱	۰/۹۸۸۸۳۶۸

برنامه NOHARM ریشه دوم میانگین مجذورات مانده‌ها (RMSR[1]) را محاسبه و به‌عنوان شاخصی برای برازش مدل ارائه می‌دهد. RMSR ریشه دوم میانگین مجذورات تفاوت بین کوواریانس‌های مشاهده شده و کوواریانس‌های پیش‌بینی شده، است. بنابراین، مقادیر کوچک RMSR نشان‌دهنده برازش مدل با داده‌ها است. یک ملاک برای تفسیر RMSR این است که با چهار برابر معکوس ریشه دوم اندازه نمونه، (خطاهای استاندارد مانده‌ها) مقایسه شود (مک دونالد[2] ۱۹۹۷). مقدار این ملاک برای تحلیل حاضر برابر ۰/۰۰۷۹۶۸۱ است. شاخص دیگر برای بررسی برازش مدل، شاخص برازش تاناکا[3](۱۹۹۳) است. به پیشنهاد مک‌دونالد (۱۹۹۷) مقدار ۰/۹۰ برای این شاخص حاکی از برازش قابل قبول و مقدار ۰/۹۵ بیانگر «برازش خوب» مدل با داده‌ها است. بنابراین، با توجه به این شاخص‌ها تک‌بعدی بودن داده‌ها پذیرفته می‌شود. در قسمت بعد پارامترهای سؤال‌ها برآورد می‌شوند.

برای بررسی اینکه کدام‌یک از مدل‌های لجستیک یک یا دوپارامتری بهتر با داده‌ها برازش دارد از آزمون خی‌دو استفاده می‌شود.

$$\chi 2 = (-2\log likelihood_{1\rho L}) - (-2\log likelihood_{2\rho L}) = 47185.8172 - 46187.2732 = 998.544$$

مقدار آماره آزمون بزرگ‌تر از مقدار بحرانی خی‌دو با ۲۹ درجه آزادی در سطح خطاپذیری ۰/۰۵ است. درنتیجه مدل دو پارامتری لجستیک بهتر از مدل لجستیک یک پارامتری با داده‌ها برازش دارد.

---

1. Root Mean square of Residuals
2. McDonald
3. Tanaka

برای برآورد پارامترهای سؤال‌ها از نرم‌افزار BILOG-MG استفاده شده است. لازم به ذکر است که مقدار آماره‌های برازش فرد H^T و l_z^* دانش‌آموزان با استفاده از بسته نرم‌افزاری PerFit موجود در نرم‌افزار R محاسبه شده است.

با استفاده از مقادیر برش آماره‌های برازش فرد، الگوی پاسخ دانش‌آموزان به دو گروه دانش‌آموزان با الگوی پاسخ مناسب و دانش‌آموزان با الگوی پاسخ نامناسب تقسیم شدند. پس از کنار گذاشتن دانش‌آموزان با الگوی پاسخ نامناسب، مجدداً پارامتر پرسش‌های آزمون با استفاده از نرم‌افزار BILOG-MG برآورد شدند. در جدول (۳) برآورد پارامتر پرسش‌ها در سه حالت کل داده‌ها، داده‌های با برازش مناسب با استفاده از آماره برازش فرد H^T، و داده‌های با برازش مناسب با استفاده از آماره برازش فرد l_z^* نمایش داده شده است. خاطرنشان می‌سازد در جدول (۳)، تغییرات بیش از ۰/۱ در برآوردهای پارامترهای پرسش‌ها مهم تلقی شده و به‌صورت توپر نمایش داده شده‌اند (سویی و موسوی، ۲۰۱۵).

جدول (۳) مقایسه برآورد پارامتر پرسش‌ها

	ضریب تشخیص			ضریب دشواری		
پرسش	کل داده‌ها	داده‌های با برازش مناسب با استفاده از آماره H^T	داده‌های با برازش مناسب با استفاده از آماره l_z^*	کل داده‌ها	داده‌های با برازش مناسب با استفاده از آماره H^T	داده‌های با برازش مناسب با استفاده از آماره l_z^*
۰۵۲۰۷۹	۰/۶۲۷	۰/۵۵۵	۰/۶۳۵	−۰/۷۰۸	**−۰/۸۶۵**	−۰/۷۳۱
۰۵۲۲۰۴	۰/۶۲۴	**۰/۵۱۸**	۰/۶۱۲	−۰/۷۵۹	**−۰/۹۸۱**	−۰/۸۰۰
۰۵۲۳۶۴	۱/۴۴۳	۱/۳۷۶	۱/۵۱۰	−۰/۵۲۷	−۰/۵۹۸	−۰/۵۳۳
۰۵۲۲۱۵	۰/۷۴۴	**۰/۶۱۴**	۰/۷۴۶	−۱/۰۲۹	**−۱,۳۳۰**	−۱/۰۹۸
۰۵۲۱۴۷	۰/۷۲۵	۰/۷۰۶	۰/۷۵۹	۰/۲۹۷	۰/۳۲۴	۰/۳۱۹
۰۵۲۰۶۷	۰/۸۶۸	**۰/۷۶۰**	۰/۸۵۲	−۰/۶۸۱	**−۰/۸۲۵**	−۰/۷۱۸
۰۵۲۰۶۸	۰/۶۹۶	۰/۶۵۳	۰/۷۳۶	۰/۹۴۵	**۱/۰۵۸**	۰/۹۷۱
۰۵۲۰۸۷	۱/۵۹۸	**۱/۴۵۱**	۱/۵۸۵	۰/۹۸۴	۱/۰۵۸	۱/۰۱۲
۰۵۲۰۴۸	۰/۵۱۶	۰/۴۱۹	۰/۵۰۱	۰/۲۱۷	**۱/۴۸۰**	۱/۲۸۵

	ضریب تشخیص			ضریب دشواری		
پرسش	کل دادهها	دادههای با برازش مناسب با استفاده از آماره H^T	دادههای با برازش مناسب با استفاده از آماره l_z^*	کل دادهها	دادههای با برازش مناسب با استفاده از آماره H^T	دادههای با برازش مناسب با استفاده از آماره l_z^*
۰۵۲۰۳۹	۱/۲۰۶	۱/۱۳۸	۱/۲۳۸	−۰/۰۶۹	−۰/۱۰۲	−۰/۰۸۹
۰۵۲۲۰۸	۱/۰۱۱	۰/۹۳۱	۱/۰۱۷	۰/۸۷۶	۰/۹۴۳	۰/۸۹۳
۰۵۲۴۱۹الف	۰/۸۴۲	۰/۷۶۶	۰/۸۷۶	−۰/۸۰۳	**−۰/۹۶۲**	−۰/۸۳۸
۰۵۲۴۱۹ب	۰/۸۱۶	۱/۷۶۸	**۲/۰۲۶**	−۱/۰۱۲	**−۱/۱۶۲**	−۱/۰۵۶
۰۵۲۱۱۵	۱/۳۲۲	۱/۲۷۷	۱/۳۶۸	−۰/۲۷۱	−۰/۳۱۷	−۰/۲۷۹
۰۵۲۴۲۱	۰/۶۲۵	۰/۶۱۱	۰/۶۴۶	۰/۱۶۰	۰/۰۹۷	۰/۱۲۹
۰۶۲۲۷۱	۱/۱۵۷	۱/۱۰۵	۱/۱۶۹	−۰/۲۶۹	−۰/۳۱۳	−۰/۲۸۳
۰۶۲۱۵۲	۱/۰۵۱	۰/۹۷۵	۱/۰۶۱	۰/۲۲۲	۰/۲۱۶	۰/۲۳۰
۰۶۲۲۱۵	۱/۱۹۶	**۱/۰۷۵**	۱/۲۱۳	۰/۹۴۰	۱/۰۲۹	۰/۹۸۱
۰۶۲۱۴۳	۱/۳۵۷	۱/۳۳۸	۱/۴۰۸	۰/۴۴۹	۰/۴۶۰	۰/۴۵۱
۰۶۲۲۳۰	۰/۲۹۵	۰/۳۱۴	۰/۳۲۴	۲/۰۹۰	**۲/۳۲۸**	۲/۱۵۶
۰۶۲۰۹۵	۰/۷۷۸	۰/۷۴۷	۰/۸۰۱	۰/۰۳۳	۰/۰۳۸	۰/۰۴۸
۰۶۲۰۷۶	۱/۰۸۱	۱/۰۲۳	۱/۰۷۲	−۰/۴۵۱	−۰/۵۱۱	−۰/۴۶۹
۰۶۲۰۳۰	۰/۴۸۶	۰/۴۴۹	۰/۴۸۷	۰/۰۲۴	−۰/۰۴۰	−۰/۰۱۱
۰۶۲۱۷۱	۱/۰۱۵	**۰/۹۰۷**	۱/۰۲۵	−۰/۷۰۸	**−۰/۸۳۷**	−۰/۷۵۰
۰۶۲۳۰۱	۰/۸۷۷	۰/۸۲۲	۰/۸۸۱	۰/۶۹۴	۰/۷۴۶	۰/۶۹۴
۰۶۲۱۹۴	۰/۸۱۷	**۰/۶۸۴**	۰/۸۰۵	−۱/۱۵۴	**−۱/۴۱۰**	−۱/۱۹۲
۰۶۲۳۴۴	۰/۸۱۲	۰/۷۵۶	۰/۸۴۲	۱/۱۶۶	**۱/۲۹۲**	۱/۲۰۳
۰۶۲۳۲۰	۱/۴۶۶	۱/۵۴۹	۱/۵۳۳	−۰/۰۱۵	−۰/۰۲۳	−۰/۰۱۰
۰۶۲۲۹۶	۱/۰۷۶	**۰/۹۷۰**	۱/۰۸۸	−۰/۵۵۱	**−۰/۶۶۹**	−۰/۵۹۶

تغییرات بیش از ۰/۱ در برآورد پارامتر پرسش‌ها به‌صورت توپر نمایش داده شده است.

با توجه به جدول (۳) برآورد ضریب تشخیص ۸ پرسش از ۲۹ پرسش (تقریباً ۲۷/۶ درصد) با استفاده از کل داده‌ها و داده‌های با برازش مناسب با استفاده از آماره H^T تغییرات بیش از ۰/۱ دارند. به همین ترتیب برآورد ضریب تشخیص ۱ پرسش از ۲۹ پرسش (تقریباً ۳/٤ درصد) با استفاده از کل داده‌ها و داده‌های با برازش مناسب با استفاده از آماره l_z^* تغییرات بیش از ۰/۱ دارند. همچنین برآورد ضریب دشواری ۱۳ پرسش از ۲۹ پرسش (٤٤/۸ درصد) با استفاده از کل داده‌ها و داده‌های با برازش مناسب با استفاده از آماره H^T تغییرات بیش از ۰/۱ دارند. به همین ترتیب برآورد ضریب دشواری هیچ‌یک از پرسش‌ها با استفاده از کل داده‌ها و داده‌های با برازش مناسب با استفاده از آماره l_z^* تغییرات بیش از ۰/۱ ندارند.

پرسش دوم: آیا تعیین الگوی پاسخ دانش‌آموزان با استفاده از آماره‌های برازش فرد H^T و l_z^* یکسان است؟

با استفاده از بسته نرم‌افزاری PerFit، برای هر یک از دانش‌آموزان مقدار آماره‌های برازش فرد H^T و l_z^* به دست آمده‌اند و الگوی پاسخ دانش‌آموزان به دو طبقه الگوی پاسخ مناسب و نامناسب، تقسیم شدند. در جدول‌های (٤ تا ٦) همسانی تعیین الگوی پاسخ دانش‌آموزان کشورهای استرالیا و ایران و جمهوری کره با استفاده از دو آماره برازش فرد H^T و l_z^* بررسی و ارائه می‌شود.

جدول (٤) همسانی تعیین الگوی پاسخ دانش‌آموزان کشور استرالیا با استفاده از آماره‌های برازش فرد H^T و l_z^*

الگوی پاسخ با استفاده از l_z^* / الگوی پاسخ با استفاده از H^T	نامناسب		مناسب		مقدار آماره آزمون	مقدار احتمال (p-value)
	تعداد	درصد	تعداد	درصد		
نامناسب	۲۷	۳/٦	۲۳	۳/۱	۱/۷۳۰	۰/۱۸۸
مناسب	۱٤	۱/۹	٦۷۷	۹۱/٤		

برای بررسی همسانی تعیین الگوی پاسخ دانش‌آموزان کشور استرالیا با استفاده از آماره‌های برازش فرد H^T و l_z^* از آزمون مک‌نمار[1] استفاده می‌شود. با توجه به جدول (٤)، مقدار آماره آزمون برابر ۱/۷۳۰ و مقدار احتمال (p-value) آن ۰/۱۸۸ است؛ بدین معنی که در سطح خطاپذیری ۰/۰۵ دلیلی برای رد همسانی تعیین الگوی پاسخ دانش‌آموزان کشور استرالیا با استفاده از آماره‌های برازش فرد H^T و l_z^* وجود ندارد.

جدول (۵) همسانی تعیین الگوی پاسخ دانش‌آموزان کشور ایران با استفاده از آماره‌های برازش فرد H^T و l_z^*

الگوی پاسخ با استفاده از l_z^* / الگوی پاسخ با استفاده از H^T	نامناسب		مناسب		مقدار آماره آزمون	مقدار احتمال (p-value)
	تعداد	درصد	تعداد	درصد		
نامناسب	۱۹	٤/٤	۱۷	٤/۰	۱/۸۸۵	۰/۱۷۰
مناسب	۹	۲/۱	۳۸۵	۸۹/۵		

برای بررسی همسانی تعیین الگوی پاسخ دانش‌آموزان کشور ایران با استفاده از آماره‌های برازش فرد H^T و l_z^* از آزمون مک‌نمار استفاده می‌شود. با توجه به جدول (۵)، مقدار آماره آزمون برابر ۱/۸۸۵ و مقدار احتمال (p-value) آن ۰/۱۷۰ است؛ بدین معنی که در سطح خطاپذیری ۰/۰۵ دلیلی برای رد همسانی تعیین الگوی پاسخ دانش‌آموزان کشور ایران با استفاده از آماره‌های برازش فرد H^T و l_z^* وجود ندارد.

[1]. McNemar Test

جدول (٦) همسانی تعیین الگوی پاسخ دانش‌آموزان جمهوری کره با استفاده از آماره‌های برازش فرد H^T و l_z^*

الگوی پاسخ با استفاده از l_z^* / الگوی پاسخ با استفاده از H^T	نامناسب		مناسب		مقدار احتمال (p-value)
	تعداد	درصد	تعداد	درصد	
نامناسب	۱۷	٤/۳	۱۱	۲/۸	۰/۳۳۲
مناسب	٦	۱/۵	۳۵۷	۹۱/۳	

برای بررسی همسانی تعیین الگوی پاسخ دانش‌آموزان جمهوری کره با استفاده از آماره‌های برازش فرد H^T و l_z^* از آزمون مک‌نمار استفاده می‌شود. با توجه به جدول (٦)، مقدار احتمال (p-value) آن ۰/۳۳۲ است؛ بدین معنی که در سطح خطاپذیری ۰/۰۵ دلیلی برای رد همسانی تعیین الگوی پاسخ دانش‌آموزان جمهوری کره با استفاده از آماره‌های برازش فرد H^T و l_z^* وجود ندارد.

پرسش سوم: آیا رابطه‌ای بین برآورد توانایی دانش‌آموزان با الگوی پاسخ آنها وجود دارد؟

در جدول‌های (۷ تا ۱۲) نتایج بررسی رابطه الگوی پاسخ دانش‌آموزان کشورهای استرالیا و ایران و جمهوری کره و برآورد توانایی آنها با استفاده از دو آماره برازش فرد H^T و l_z^* ارائه می‌شود.

جدول (۷) رابطه الگوی پاسخ دانش‌آموزان کشور استرالیا و برآورد توانایی آنها با استفاده از آماره برازش فرد H^T

برآورد توانایی	نامناسب		مناسب		مقدار آماره آزمون	مقدار احتمال (p-value)
	تعداد	درصد	تعداد	درصد		
کمتر از ۱-	۳۰	۳۰/۳	٦۹	٦۹/۷	۱۰۰/۷۸۳	<۰/۰۰۱
بین ۱- و ۱	۱۷	۳/۲	۵۱۷	۹٦/۸		
بیشتر از ۱	۳	۲/۸	۱۰۵	۹۷/۲		

برای بررسی رابطه سطوح توانایی دانش‌آموزان کشور استرالیا و الگوی پاسخ آنها (مناسب یا نامناسب) با استفاده از آماره برازش فرد H^T از آزمون خی‌دو استفاده می‌شود. با توجه به جدول (۷)، مقدار آماره آزمون برابر ۱۰۰/۷۸۳ و مقدار احتمال (p-value) آن کمتر از ۰/۰۰۱ است؛ بدین معنی که در سطح خطاپذیری ۰/۰۵ استقلال بین سطوح توانایی دانش‌آموزان کشور استرالیا و الگوی پاسخ آنها رد می‌شود.

در جدول (۸)، نتایج بررسی رابطه سطوح توانایی دانش‌آموزان ایران و الگوی پاسخ آنها با استفاده از آماره برازش فرد H^T ارائه می‌شود.

جدول (۸) رابطه الگوی پاسخ دانش‌آموزان ایران و توانایی آنها با استفاده از آماره برازش فرد H^T

	نامناسب		مناسب		مقدار آماره آزمون	مقدار احتمال (p-value)
توانایی	تعداد	درصد	تعداد	درصد		
کمتر از ۱-	۱۳	۲۸/۹	۳۲	۷۱/۱	۲۹/۷۶۲	<۰/۰۰۱
بین ۱- و ۱	۲۲	۶/۹	۲۹۵	۹۳/۱		
بیشتر از ۱	۱	۱/۵	۶۷	۹۸/۵		

برای بررسی رابطه سطوح توانایی دانش‌آموزان ایران و الگوی پاسخ آنها (مناسب یا نامناسب) با استفاده از آماره برازش فرد H^T از آزمون خی‌دو استفاده شده است. با توجه به جدول (۸) مقدار آماره آزمون برابر ۲۹/۷۶۲ و مقدار احتمال (p-value) آن کمتر از ۰/۰۰۱ است؛ بدین معنی که در سطح خطاپذیری ۰/۰۵ استقلال بین سطوح توانایی دانش‌آموزان کشور ایران و الگوی پاسخ آنها رد می‌شود.

در جدول (۹)، نتایج بررسی رابطه سطوح توانایی دانش‌آموزان جمهوری کره و الگوی پاسخ آنها با استفاده از آماره برازش فرد H^T ارائه می‌شود.

جدول (۹) رابطه الگوی پاسخ دانش‌آموزان جمهوری کره و توانایی آنها با استفاده از آماره برازش فرد H^T

	نامناسب		مناسب		مقدار آماره آزمون	مقدار احتمال (p-value)
توانایی	تعداد	درصد	تعداد	درصد		
کمتر از ۱-	۳	٤/۷	٦۱	۹٥/۳	۲۸/۹۱۷	<۰/۰۰۱
بین ۱- و ۱	۱۱	۳/۹	۲۷۱	۹٦/۱		
بیشتر از ۱	۱٤	۳۱/۱	۳۱	٦۸/۹		

برای بررسی رابطه سطوح توانایی دانش‌آموزان جمهوری کره و الگوی پاسخ آنها (مناسب یا نامناسب) با استفاده از آماره برازش فرد H^T به‌دلیل اینکه فراوانی مورد انتظار ۲ خانه از ٦ خانه جدول (۳۳/۳ درصد) کمتر از ٥ است از آزمون دقیق فیشر[1] استفاده شده است. با توجه به جدول (۹)، مقدار آماره آزمون برابر ۲۸/۹۱۷ و مقدار احتمال (p-value) آن کمتر از ۰/۰۰۱ است؛ بدین معنی که در سطح خطاپذیری ۰/۰٥ استقلال بین سطوح توانایی دانش‌آموزان جمهوری کره و الگوی پاسخ آنها رد می‌شود.

در جدول (۱۰)، نتایج بررسی رابطه سطوح توانایی دانش‌آموزان کشور استرالیا و الگوی پاسخ آنها با استفاده از آماره برازش فرد l_z^* ارائه می‌شود.

جدول (۱۰) رابطه الگوی پاسخ دانش‌آموزان کشور استرالیا و توانایی آنها با استفاده از آماره برازش فرد l_z^*

	نامناسب		مناسب		مقدار آماره آزمون	مقدار احتمال (p-value)
توانایی	تعداد	درصد	تعداد	درصد		
کمتر از ۱-	۱۰	۹/٥	۹٥	۹۰/٥	۳/۹٦۷	۰/۱۳۸
بین ۱- و ۱	۲٥	٤/۷	٥۰۹	۹٥/۳		
بیشتر از ۱	٦	٥/٦	۱۰۲	۹٤/٤		

[1]. Fisher's Exact Test

برای بررسی رابطه سطوح توانایی دانش‌آموزان کشور استرالیا و الگوی پاسخ آنها (مناسب یا نامناسب) با استفاده از آماره برازش فرد l_z^* از آزمون خی‌دو استفاده شده است. با توجه به جدول (۱۰)، مقدار آماره آزمون برابر ۳/۹٦۷ و مقدار احتمال (p-value) آن ۰/۱۳۸ است؛ بدین معنی که در سطح خطاپذیری ۰/۰۵ دلیلی برای رد استقلال بین سطوح توانایی دانش‌آموزان کشور استرالیا و الگوی پاسخ آنها وجود ندارد.

در جدول (۱۱)، نتایج بررسی رابطه سطوح توانایی دانش‌آموزان ایران و الگوی پاسخ آنها با استفاده از آماره برازش فرد l_z^* ارائه می‌شود.

جدول (۱۱) رابطه الگو پاسخ دانش‌آموزان ایران و توانایی آنها با استفاده از آماره برازش فرد l_z^*

	نامناسب		مناسب		مقدار آماره	مقدار احتمال
توانایی	تعداد	درصد	تعداد	درصد	آزمون	(p-value)
کمتر از ۱-	۳	٦/۵	٤۳	۹۳/۵	۱/۵۳۷	۰/٤۹٤
بین ۱- و ۱	۲۳	۷/۳	۲۹٤	۹۲/۷		
بیشتر از ۱	۲	۲/۹	٦٦	۹۷/۱		

برای بررسی رابطه سطوح توانایی دانش‌آموزان ایران و الگوی پاسخ آنها (مناسب یا نامناسب) با استفاده از آماره برازش فرد l_z^* به‌دلیل اینکه فراوانی مورد انتظار ۲ خانه از ٦ خانه جدول (۳۳/۳ درصد) کمتر از ۵ است، از آزمون دقیق فیشر استفاده شده است. با توجه به جدول (۱۱)، مقدار آماره آزمون برابر ۱/۵۳۷ و مقدار احتمال (p-value) آن ۰/٤۹٤ است؛ بدین معنی که در سطح خطاپذیری ۰/۰۵ دلیلی برای رد استقلال بین سطوح توانایی دانش‌آموزان ایران و الگوی پاسخ آنها وجود ندارد.

در جدول (۱۲)، نتایج بررسی رابطه سطوح توانایی دانش‌آموزان جمهوری کره و الگوی پاسخ آنها با استفاده از آماره برازش فرد l_z^* ارائه می‌شود.

جدول (۱۲) رابطه الگوی پاسخ دانش‌آموزان جمهوری کره و توانایی آنها با استفاده از آماره برازش فرد l_z^*

	نامناسب		مناسب		مقدار آماره آزمون	مقدار احتمال (p-value)
توانایی	تعداد	درصد	تعداد	درصد		
کمتر از ۱-	۳	٤/۷	٦۱	۹٥/۳	۰/۱۱۳	۱/۰۰
بین ۱- و ۱	۱۷	٦/۰	۲٦٥	۹٤/۰		
بیشتر از ۱	۳	٥/٥	٥۲	۹٤/٥		

برای بررسی رابطه سطوح توانایی دانش‌آموزان جمهوری کره و الگوی پاسخ آنها (مناسب یا نامناسب) با استفاده از آماره برازش فرد l_z^* به‌دلیل اینکه فراوانی مورد انتظار ۲ خانه از ٦ خانه جدول (۳۳/۳ درصد) کمتر از ٥ است، از آزمون دقیق فیشر استفاده شده است. با توجه به جدول (۱۲)، مقدار آماره آزمون برابر ۰/۱۱۳ و مقدار احتمال (p-value) آن ۱/۰۰ است؛ بدین معنی که در سطح خطاپذیری ۰/۰٥ دلیلی برای رد استقلال بین سطوح توانایی دانش‌آموزان جمهوری کره و الگوی پاسخ آنها وجود ندارد.

بحث و نتیجه‌گیری

در این پژوهش از آماره‌های برازش فرد H^T و l_z^* برای بررسی الگوی پاسخ آزمون ریاضی تیمز ۲۰۱٥ دانش‌آموزان پایه هشتم کشورهای استرالیا، ایران و جمهوری کره استفاده شده است. ابتدا مفروضه تک‌بعدی بودن بررسی شد. با تأیید این مفروضه به برازش نسبی مدل متناسب با داده‌ها اقدام شد، بهترین مدل برازش داده شده با داده‌ها، مدل دوپارامتری لجستیک بود. سپس به برآورد پارامتر پرسش‌ها اقدام شد. از طرفی با محاسبه آماره‌های برازش فرد H^T و l_z^* هر یک از دانش‌آموزان، الگوی پاسخ آنها به دو طبقه الگوی پاسخ با برازش مناسب و الگوی پاسخ با برازش نامناسب تعیین شد. با حذف دانش‌آموزان با الگوی پاسخ با برازش نامناسب توسط هر یک از آماره‌های برازش فرد H^T و l_z^*، مجدداً پارامتر پرسش‌ها برآورد شد. با مقایسه برآورد پارامتر پرسش‌ها در وضعیت‌های مذکور مشخص شد وجود الگوی پاسخ نامناسب بر برآورد پارامتر پرسش‌ها تأثیرگذار است. در این باره، نتایج این پژوهش با یافته‌های

پژوهش‌های فیلیپس (۱۹۸۶)، سوتاریدنا و همکاران (۲۰۰۵)، هندراوان و همکاران (۲۰۰۵)، و موسوی (۲۰۱۵) همسویی دارد.

همچنین برای بررسی همسانی تعیین الگوی پاسخ دانش‌آموزان توسط دو آماره برازش فرد H^T و l_z^* در کشورهای استرالیا، ایران و جمهوری کره مشخص شد که برای هر سه کشور دلیلی برای رد فرض همسانی تعیین الگوی پاسخ دانش‌آموزان توسط دو آماره برازش فرد H^T و l_z^* وجود ندارد. لازم به ذکر است این پژوهش از معدود پژوهش‌هایی است که با استفاده از داده‌های تجربی، همسانی تعیین الگوی پاسخ افراد با آماره‌های برازش فرد متفاوت بررسی شده است.

در بررسی برآورد توانایی دانش‌آموزان و الگوی پاسخ آنها مشخص شد در کشور استرالیا با استفاده از آماره برازش فرد H^T، رابطه معنی‌داری بین سطوح برآورد توانایی دانش‌آموزان و الگوی پاسخ آنها وجود دارد به‌طوری که با افزایش سطوح برآورد توانایی دانش‌آموزان، از درصد دانش‌آموزان با الگوی پاسخ نامناسب کاسته و به درصد دانش‌آموزان با الگوی پاسخ مناسب افزوده شده است. در همین ارتباط و در ایران نیز با استفاده از آماره برازش فرد H^T، رابطه معنی‌داری بین سطوح برآورد توانایی دانش‌آموزان و الگوی پاسخ آنها وجود دارد به‌طوری که با افزایش سطوح برآورد توانایی دانش‌آموزان، از درصد دانش‌آموزان با الگوی پاسخ نامناسب کاسته و به درصد دانش‌آموزان با الگوی پاسخ مناسب افزوده شده است. در همین ارتباط در جمهوری کره نیز با استفاده از آماره برازش فرد H^T، رابطه معنی‌داری بین سطوح برآورد توانایی دانش‌آموزان و الگوی پاسخ آنها وجود دارد ولیکن برخلاف دو کشور استرالیا و ایران، با افزایش سطوح برآورد توانایی دانش‌آموزان از طبقه کمتر از یک به طبقه بیشتر از یک، از درصد دانش‌آموزان با الگوی پاسخ مناسب کاسته و به درصد دانش‌آموزان با الگوی پاسخ نامناسب افزوده شده است. در ارتباط با آماره برازش فرد l_z^*، بین سطوح برآورد توانایی دانش‌آموزان و الگوی پاسخ آنها در هیچ‌یک از کشورهای استرالیا، ایران و جمهوری کره رابطه معنی‌داری وجود نداشت.

همچنین در بررسی برآورد توانایی دانش‌آموزان و الگوی پاسخ آنها با استفاده از آماره‌های برازش فرد H^T و l_z^* مشخص شد که برای هر سه کشور استرالیا، ایران و جمهوری کره رابطه معنی‌داری بین مقدار آماره برازش فرد H^T دانش‌آموزان و برآورد توانایی آنها وجود دارد، در حالی که همین رابطه برای آماره برازش فرد l_z^*، معنی‌دار نیست. این عدم همسانی نتایج بین دو آماره برازش فرد H^T و l_z^* را می‌توان

بدین دلیل دانست که در فرمول آماره برازش فرد l_z^*، برآورد توانایی دانش‌آموزان لحاظ شده در حالی که آماره برازش فرد H^T، از نوع آماره‌های برازش فرد ناپارامتریک است.

آماره‌های برازش فرد چون پاسخگوی این نیست که برازش نامناسب چگونه رخ می‌دهد، مورد انتقاد قرار گرفته است. با این حال ما معتقد هستیم که سنجش برازش فرد برای توسعه و اعتبارسنجی آزمون‌ها مفید است. آماره‌های برازش فرد را می‌توان در مطالعه مقدماتی آزمون‌ها مورد استفاده قرار داد. اگر در بررسی نتایج مطالعه مقدماتی آزمون مشخص شد که درصد الگوهای پاسخ با برازش نامناسب بیش از حد مورد انتظار بود، لازم است به بررسی بیشتر محتوای آزمون‌ها اقدام شود. از این‌رو، سنجش برازش فرد توان بالقوه‌ای برای کمک به بهبود روایی آزمون دارد و با شناسایی مشکلات احتمالی پیش روی آزمون و برطرف کردن آنها، می‌توان به آزمون‌هایی عادلانه‌تر و دقیق‌تر دست یافت. تجزیه‌وتحلیل تعقیبی برای بررسی دقیق‌تر دلایل ممکن برازش نامناسب، باید اجرا شود.

پاسخ‌های با برازش نامناسب را که صرفاً بر اساس آماره‌های برازش فرد شناسایی شده‌اند نباید هرگز به‌عنوان شواهد کافی برای نامعتبر شناختن نتایج آزمون دانشجویان لحاظ کرد. تجزیه‌وتحلیل تعقیبی برای بررسی دقیق‌تر دلایل ممکن برازش نامناسب باید اجرا شود. اگر فرضیه خاصی در ارتباط با نوع رفتار با برازش نامناسب دانش‌آموزان وجود داشته باشد، روش‌هایی از قبیل اندازه‌گیری مناسب بهینه برای آزمون آماری آن فرضیه در دسترس است (لوین و دراسگو، ۱۹۸۸؛ لامپریانو[1] ۲۰۱۰). مزیت چنین آزمون فرضیه‌هایی این است که نتایج آنها در مورد نوع برازش نامناسب آگاهی‌بخش‌تر است. با این حال، برای درک واقعی اینکه چرا برازش نامناسب رخ داده است، اطلاعاتی درباره پاسخ دانش‌آموزان از قبیل مصاحبه با دانش‌آموزان و معلمان، گزارش‌های کلامی دانش‌آموزان، اطلاعات ردیابی چشمی، زمان واکنش ممکن است مورد نیاز باشد (انجمن تحقیقات آموزشی آمریکا[2] شورای ملی اندازه‌گیری در آموزش‌وپرورش[3] و انجمن روان‌شناسی آمریکا[4] ۱۹۹۹). این نوع از

---

[1]. Lamprianou
[2]. American Educational Research Association
[3]. National Council for Measurement in Education
[4]. American Psychological Association

اطلاعات تصویر نسبتاً مشروحی از چگونگی پاسخ به پرسش‌های آزمون‌ها فراهم می‌کند که می‌تواند به شناسایی عوامل مرتبط با برازش نامناسب کمک کند. بدین ترتیب نتایج حاصل از آماره‌های برازش فرد به‌طور قابل ملاحظه‌ای قابل تفسیر و ارزشمند هستند و مهم‌تر از همه اینکه با استفاده از آماره‌های برازش فرد می‌توان روایی نمره‌های آزمون را بهبود بخشید.

تشکر و قدردانی

در اینجا لازم است از دکتر مسعود کبیری، مدیر فنی و داده‌پردازی مرکز ملی مطالعات تیمز و پرلز پژوهشگاه مطالعات آموزش‌وپرورش، به‌خاطر کمک‌های متعددشان در طول پژوهش، تشکر و قدردانی شود.

منابع

سرمد، زهره؛ بازرگان، عباس و حجازی، الهه (۱۳۸٤). *روش‌های تحقیق در علوم رفتاری*. تهران: نشر آگاه.

کبیری، مسعود؛ کریمی، عبدالعظیم و بخشعلی‌زاده، شهرناز (۱۳۹۵). *یافته‌های ملی تیمز ۲۰۱۵، روند ۲۰ ساله آموزش علوم و ریاضیات ایران در چشم‌انداز بین‌المللی*. پژوهشگاه مطالعات آموزش‌وپرورش، انتشارات مدرسه.

مینایی، اصغر و فلسفی‌نژاد، محمدرضا (۱۳۸۹). روش‌های سنجش تک‌بعدی بودن سؤال‌ها در مدل‌های دوارزشی IRT. *فصلنامه اندازه‌گیری تربیتی، ۱* (۳)، ۷۱–۱۰۰.

Albers, C. J.; Meijer, R. R. & Tendeiro, J. N. (2016). Derivation and applicability of asymptotic results for multiple subtests person-fit statistics. *Applied Psychological Measurement, 40* (4), 274-288.

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

Childs, R. A. & Jaciw, A. P. (2003). Matrix sampling of items in large-scale assessments. *Practical Assessment, Research & Evaluation, 8* (16), 1 – 9.

Conijn, J. M.; Emons, W. H. M. & Sijtsma, K. (2014). Statistics l_z-based person-fit methods for noncognitive multiscale measures. *Applied Psychological Measurement, 38* (2), 122-136.

Cui, Y. & Li, J. (2015). Evaluating Person fit for cognitive diagnostic assessment. *Applied Psychological Measurement, 39* (3), 223-238.

Cui, Y. & Mousavi, A. (2015). Explore the usefulness of person-fit analysis on large-scale assessment. *International Journal of Testing, 15* (1), 23-49.

De Champlain, A. F. & Gessaroli, M. F. (1998). Assessing the dimensionality of item response matrices with small sample size and short test lengths. *Applied Measurement in Education, 11* (1), 231-253.

Drasgow, F.; Levine, M. V. & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38* (1), 67–86.

Finch, H. & Habing, B. (2005). Comparison of NOHARM and DETECT in item cluster recovery: Counting dimensions and

allocating items. *Journal of Educational Measurement, 42* (2), 149-170.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9*, 139–150.

Harnisch, D. L. & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18* (3), 133–146.

Hendrawan, I.; Glas, C. A. & Meijer, R. R. (2005). The effect of person misfit on classification decisions. *Applied psychological measurement, 29* (1), 26-44.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics, *Applied Measurement in Education, 16* (4), 277-298.

Knol, D. L. & Berger, P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26* (3), 457-477.

Lamprianou, I. & Boyle, B. (2004). Accuracy of measurement in the context of mathematics national curriculum tests in England for ethnic minority pupils and pupils who speak English as an additional language. *Journal of Educational Measurement,* 41 (3), 239–259.

Levine, M. V. & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical & Statistical Psychology, 35* (1), 42–56.

Levine, M. V. & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika, 53* (2), 161–176.

Magis, D.; Raiche, G. & Beland, S. (2012). A didactic presentation of Snijder's l_z^* index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational & Behavioral Statistics, 37* (1), 57-81.

Martin, M. O.; Mullis, I. V. S. & Hooper, M. (2016). *Methods and procedures in TIMSS 2015*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

McDonald, R. P. (1997). Normal-ogive multidimensional model. In W. J. van der Linden & R. K. Hambleton (Ed.), *Handbook of Modern Item Response Theory* (pp. 258-269). New York: Springer Verlag.

Meijer, R. R. (1997). Person fit and criterion-related validity: An extension of the Schmitt, Cortina, and Whitney study. *Applied Psychological Measurement, 21* (2), 99 -113.

Meijer, R. R. & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25* (2), 107-135.

Mousavi, S. A. (2015). *The effect of person misfit on item parameter estimation: A simulation study*. Doctoral dissertation, University of Alberta.

Mousavi, A. Tendeiro, J. N. & Younesi, J. (2016). Person fit assessment using the PerFit package in R. *The Quantitative Methods for Psychology, 12* (3), 232-242.

Olson, J. F. Martin, M. O. & Mullis, I. V.S. (2008). *TIMSS 2007 Technical Report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Phillips, S. E. (1986). The effects of deletion of misfitting persons on vertical equating via the Rasch model. *Journal of Educational Measurement, 23* (2), 107–118.

Rudner, L. M. Bracey, G. & Skaggs, G. (1996). The use of a person-fit statistic with one high quality achievement test. *Applied Measurement in Education, 9* (1), 91–109.

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in Item Response Theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test & Assessment Modeling, 55* (1), 3-38.

Schmitt, N. S. Cortina, J. M. & Whitney, D. J. (1993). Appropriateness fit and criterion-related validity. *Applied Psychological Measurement, 17* (2), 143-150.

Sijtsma, K. (1986). A coefficient of deviance of response patterns. *Kwantitatieve Methoden, 7*(22), 131–145.

Sijtsma, K. & Meijer, R. R. (1992). A method for investigating the intersection of item response function in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16* (2), 149-157.

Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational & Psychological Measurement, 45* (3), 433–444.

Smith, R. M. (1986). Person fit in the Rasch model. *Educational & Psychological Measurement, 46* (2), 359–372.

Snijders, T. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika, 66* (3), 331-342.

Sotaridona, L. S.; Choi, S. W. & Meijer, R. R. (2005). *The Effect of Misfitting Response Vectors on Item Calibration and Performance Classification*. Retrieved May 2013, from CTB/McGraw-Hill: http://www.ctb.com/img/pdfs/raMisfittingResponseVectors.pdf

Stocking, M. L. & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7* (2), 201-210.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49* (1), 95–110.

Tatsuoka, K. K. & Tatsuoka, M. M. (1983). Spotting erroneous rules of operation by the individual consistency index. *Journal of Educational Measurement, 20* (3), 221–230.

Van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology, 13* (3), 267–298.

Wright, B. D. & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D. & Stone, M. H. (1979). *Best test design. Rasch measurement*. Chicago: Mesa Press.