فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی
سال نهم، شماره ۲۵، بهار ۱۳۹۸، صص ۲۹-۷

Journal of Educational Measurement & Evaluation Studies
Vol. 9, No. 25, Spring 2019

تاریخ دریافت مقاله: ۹۶/۰۸/۱۳

تاریخ پذیرش مقاله: ۹۶/۱۲/۲۱

افت پارامترهای سؤال‌های روند آزمون ریاضیات پایه هشتم تیمز ایران و تأثیر آن در برآورد پارامتر توانایی[۱]

اصغر مینائی*
ابوالفضل قدمی**
نورعلی فرخی***

چکیده

هدف از اجرای این پژوهش بررسی افت پارامترهای سؤال (IPD) روند آزمون ریاضیات تیمز پایه هشتم ایران و تأثیر آن بر برآورد پارامتر توانایی دانش‌آموزان بود. جامعه آماری شامل همه دانش‌آموزان پایه هشتم شرکت‌کننده در آزمون ریاضیات تیمز ایران در سال‌های ۲۰۰۳، ۲۰۰۷ و ۲۰۱۱ بود؛ ۱۲۲٤ نفر (٤۰۸ نفر در هر دوره) از دانش‌آموزانی که به ٤۰ سؤال روند موجود در این آزمون پاسخ داده بودند. برای بررسی وجود یا نبود افت از نرم‌افزار $SPSS_{۲۳}$ و روش رگرسیون لجستیک استفاده شد. بر اساس نتایج از بین ٤۰ سؤال روند ۲ سؤال افت را در پارامترهای خود نشان دادند؛ که هر دو سؤال نیز دارای افت پارامتر یکنواخت بود؛ اما طبق ملاک اندازه اثر جودوین و جیرل هر دو سؤال اندازه اثر ناچیزی داشتند. برای بررسی تأثیر افت پارامترها بر توانایی دانش‌آموزان، ابتدا توانایی (θ) دانش دانش‌آموزان با حضور کل سؤال‌ها محاسبه شد و پس از شناسایی و حذف سؤال‌های دارای افت، بار دیگر پارامتر توانایی (θ) با استفاده از نرم‌افزار BILOG-MG محاسبه شد؛ سپس برای بررسی میزان تفاوت میانگین‌ها و معنی‌داری آن از روش آماری t وابسته استفاده شد؛ نتیجه آنکه این میزان افت در برآورد توانایی دانش‌آموزان تأثیر چشمگیری نداشته است.

واژگان کلیدی: افت پارامترهای سؤال، تیمز، رگرسیون لجستیک، تغییرناپذیری

---

۱. این مقاله از پایان نامه کارشناسی ارشد نویسنده دوم استخراج شده است.

* استادیار گروه سنجش و اندازه‌گیری، دانشگاه علامه طباطبایی

**کارشناس ارشد سنجش و اندازه‌گیری، دانشگاه علامه طباطبایی(نویسنده مسئول ghadami63pnu@gmail.com)

*** دانشیار گروه سنجش و اندازه‌گیری، دانشگاه علامه طباطبایی

مقدمه

آزمون‌های بین‌المللی تیمز[1] از مهم‌ترین مطالعات در زمینه ارزشیابی پیشرفت تحصیلی ریاضیات و علوم است که بیش از ٦٠ کشور از سراسر جهان در آن عضویت دارند (مولیس، مارتین، فوی و اورا[2]، ۲۰۱۲). آزمون‌های بین‌المللی این امکان را فراهم می‌آورند تا پیشرفت تحصیلی کشورهای مختلف با یکدیگر مقایسه شود و چشم‌انداز وسیعی را برای ارزیابی و بهبود آموزش فراهم می‌آورد (ارسی کان[3]، ۱۹۹۸، به نقل از مینائی، ۱۳۹۲). معادل یا هم‌ارز بودن سازه در میان گروه‌های مورد مطالعه شرط اصلی برای بررسی این تفاوت‌های گروهی است که از آن با عنوان مقایسه‌پذیری سازه نام برده می‌شود و میزان روایی که از سنجش‌های بین‌المللی استنباط می‌شود به میزان مقایسه‌پذیری آزمون در میان کشورها بستگی دارد (مینائی، ۱۳۹۲).

نظام آموزشی کشور، مرکزی برای سنجش ملی پیشرفت تحصیلی منسجم و سازمان‌یافته جهت سنجش و ارزشیابی بازده‌های آموزشی ندارد. در نبود چنین مرکزی، شرکت در مطالعات انجمن بین‌المللی ارزشیابی پیشرفت تحصیلی با عنوان «مطالعه بین‌المللی روند تغییرات ریاضی و علوم» فرصت مناسبی برای بررسی عملکرد نظام آموزشی در سطح ملی و مقایسه آن با عملکرد سایر کشورها فراهم آورده است؛ به عبارتی، در شرایط فعلی تنها از طریق یافته‌های حاصل از شرکت در مطالعات بین‌المللی می‌توان عملکردها و برون‌دادهای نظام آموزشی را ارزشیابی و هم‌زمان با سایر کشورها مقایسه کرد (کیامنش، صفرخانی، کبیری و اقدسی، ۱۳۹۲).

از زمان برگزاری آزمون‌های تیمز (۱۹۹۵) تاکنون هزینه‌های بسیاری صرف اجرای آزمون‌های تیمز در ایران شده است. کسب نتایج ضعیف و قرار گرفتن دانش‌آموزان ایرانی در رتبه‌ای پایین‌تر از میانگین جهانی در دوره‌های برگزار شده (مولیس و همکاران، ۲۰۱۲؛ مولیس، مارتین، گونزالس و کروستوسکی[4]، ۲۰۰٤) اهمیت این موضوع را مشخص می‌سازد که برای شناخت علل این عدم موفقیت‌ها پژوهش‌هایی در حوزه‌های مختلف صورت گیرد. پژوهش‌های بسیاری نیز در حوزه‌های مختلف

---

1. TIMSS
2. Mullis, Martin, Foy & Arora
3. Ercikan
4. Mullis, Martin, Gonzalez & Chrostowski

این آزمون صورت گرفته است (مینائی،۱۳۹۱ و ۱۳۹۲؛ کبیری،۱۳۹۲؛ میرزاخانی و فرزاد، ۱۳۹۲)؛ اما افت پارامترهای سؤال[1] (IPD) در مطالعات سنجش و اندازه‌گیری توجه چندانی به خود جلب نکرده و مطالعات اندکی در این زمینه صورت گرفته است.

افت پارامتر دلالت بر اختلاف یا تغییری در پارامترهای سؤال آزمون «درجه دشواری، قدرت تشخیص» در دوره‌های مختلف زمانی دارد و این موضوع هنگامی اتفاق می‌افتد که یک آزمون در طی دوره‌های مختلف اجرا شود ولی پارامترهای سؤال برای آزمودنی‌ها به دلایل مختلف تغییر کند (گلدشتاین[2]، ۱۹۸۳). به نظر گلدشتاین (۱۹۸۳) مطالعه و بررسی افت پارامترهای سؤال در آزمون‌های روند[3]، به متخصصان آزمون ساز کمک می‌کند تا بدانند یک سؤال تا چه زمان پارامترهای اولیه خود را حفظ می‌کند و قابلیت اجرا دارد و چه موقع آزمون باید مورد تجدیدنظر قرار گیرد تا دوره‌های مختلف قابل مقایسه باشند و موجب تصمیم‌گیری غیرمنصفانه نشود (پارک، لی و زینگ[4]، ۲۰۱۶).

باک، موراک و پفیفنبرگر[5] (۱۹۸۸) تکرار بیش‌ازحد سؤال‌های موجود در بانک سؤال و موقعیت‌های مختلف که در آن شرایط آزمون گرفته می‌شود را خطری جدی و عامل ایجادکننده افت پارامترهای سؤال می‌دانند. کولن و برنان[6] (۱۹۹۵) دلایلی که در افت پارامتر سؤال می‌توانند دخیل باشند را نقص در سنجش اولیه و پارامترپردازی سؤال‌ها، جایگاه سؤال در فرم‌های مختلف، تغییر در طراحی برگه پاسخنامه، تغییر دادن جایگاه سؤال در پرسشنامه، تغییر در فونت یا صفحه‌بندی برگه امتحانی و برگزاری امتحان در شرایط غیراستاندارد نام برده‌اند. همچنین نگهداری بلندمدت سؤال‌ها در بانک سؤال نیز به سبب استفاده بیش‌ازحد سؤال‌ها و آشنایی پاسخ‌دهندگان

---

1. Item Parameter Drift
2. Goldstein
3. Trend
4. Park, Lee & Xing
5. Bock, Murake & Pfeiffenberger
6. Kolen & Brennan

با آن می‌شود، درنتیجه درجه دشواری سؤال‌ها ممکن است کاهش پیدا کند و روایی آزمون را تحت تأثیر خود قرار دهد (وندرلیندن و گلس[1]، ۲۰۱۰).

رایج‌ترین نوع افت در پارامتر سؤال‌ها تغییر در پارامتر دشواری است (باک و همکاران، ۱۹۸۸). این نوع افت به دلایل مختلفی از قبیل تغییر در دانش و اطلاعات دانش‌آموزان در طول زمان، فاش‌سازی امتحان به دلیل تقلب، تغییر در برنامه درسی، تمرین و سیاست‌های آموزشی، وقایع تاریخی و مسائل فنی مربوط به آزمون به وجود می‌آید (ولز، همبلتون، کرک‌پاتریک و منج[2]، ۲۰۱٤)؛ خلاصه‌ای از دلایلی که باعث دشوارتر و آسان‌تر شدن سؤال‌ها می‌شود در جدول (۱) ارائه شده است.

جدول (۱) دلایل تغییرات در درجه دشواری سؤال‌ها

دلایل سخت‌تر شدن سؤال	دلایل آسان‌تر شدن سؤال
تغییر در شیوه اجرا، سیاست، برنامه درسی	تکرار بیش‌ازحد
تغییرات فرهنگی	فاش شدن سؤال به دلیل تقلب
وقایع تاریخی	پارامترپردازی ضعیف اولیه
موقعیت قرار گرفتن سؤال	وقایع تاریخی
پارامترپردازی ضعیف اولیه	موقعیت قرار گرفتن سؤال
	تغییرات فرهنگی

* اقتباس شده از ماکاس ریسک[3] (۲۰۱۵)

هنگامی‌که یک سؤال در طول زمان ازنظر پاسخگویی دشوارتر می‌شود، می‌تواند به دلایل مختلفی از قبیل آموزشی، فنی و تغییرات فرهنگی باشد. سنجش ضعیف اولیه سؤال‌ها می‌تواند سبب تغییرات در پارامتر دشواری سؤال، چه از نوع آسان‌تر شدن و یا سخت‌تر شدن سؤال‌ها شود. گاهی اوقات اصطلاحات فنی در طی زمان منسوخ می‌شوند و درنتیجه کاربرد اولیه خود را از دست می‌دهند. علاوه بر این، تغییر مکان سؤال در آزمون (جابجایی مکان سؤال) نیز می‌تواند باعث دشوارتر یا آسان‌تر شدن

1. Van Der Linden & Glas
2. Wells, Hambleton, Kirkpatrick & Meng
3. Risk, N. M.

سؤال در هر دو موقعیت شود؛ مانند هنگامی‌که سؤال در پیش‌آزمون در یک مکان و در آزمون اصلی در مکان دیگر قرار داشته باشد (داناهو و ایشام[۱]، ۱۹۹۸)؛ زیرا طبق تعریف، از جمله مواردی که پایایی آزمون را تحت تأثیر خود قرار می‌دهد یکسان نبودن شرایط آزمون است (سیف، ۱۳۹۰)؛ بنابراین زمانی که یک آزمون به هر دلیلی شرایط متفاوتی را برای آزمودنی فراهم کند اعتبار آزمون را خدشه‌دار می‌کند و سبب نتایج متفاوتی در آن می‌شود.

وجود افت پارامتر در سؤال می‌تواند باعث تخطی از ویژگی نامتغیر بودن[۲] پارامترهای سؤال شود و خطر جدی برای سنجش و اندازه‌گیری محسوب شود. افت پارامتر سؤال می‌تواند خطاهای اندازه‌گیری را بیش از اندازه بزرگ کند و روی روایی محتوا و سازه یک سؤال تأثیرگذار باشند (مک‌کوی[۳]، ۲۰۰۹). وجود افت همچنین روی روش‌های مبتنی بر نظریه جدید اندازه‌گیری[٤] مانند هم‌ارز سازی[٥] و آزمون‌های تطبیقی[٦] تأثیرگذار است و می‌تواند باعث ایجاد خطای هم‌ارز سازی شود؛ علاوه بر آنچه تاکنون در این مورد گفته شد، افت پارامتر برای سؤال‌هایی که به ثبات در طول زمان نیاز دارد نیز خطری جدی محسوب می‌شود (ولز، سابکوویاک و سرلین[۷]، ۲۰۰۲؛ ولز، همبلتون، کرک‌پاتریک و منج، ۲۰۱٤)؛ این ثبات در آزمون‌هایی که با هدف مقایسه نمره‌های افراد در دوره‌های مختلف اجرا می‌شوند از اهمیت ویژه‌ای برخوردار است (گوا و وانگ[۸]، ۲۰۰۳).

برای بررسی میزان افت پارامترهای سؤال از روش‌هایی همانند کارکرد افتراقی استفاده می‌شود. با این تفاوت که به جای بررسی گروه‌های مختلف، دوره‌های مختلف زمانی بررسی می‌شود. روش‌های آماری برای شناسایی افت پارامترها به دو دسته کلی مبتنی بر نظریه کلاسیک و نظریه‌های جدید اندازه‌گیری تقسیم می‌شوند. روش‌های

---

1. Donoghue & Isham
2. Invariance
3. McCoy
4. Item Response Theory
5. equating
6. Comparative tests
7. Wells, Subkoviak & Serlin
8. Guo & Wang

مبتنی بر نظریه کلاسیک، نمره‌های مشاهده شده را به‌عنوان متغیری همتا برای پیش‌بینی نمره‌های واقعی در نظر می‌گیرند (لرد و ناویک[1]، ۱۹٦۸). از مزیت و برتری‌های روش‌های مبتنی بر نظریه کلاسیک، سادگی اجرا و کار با نمونه‌های با حجم کوچک است (امبرتسون، رایس و استیون، ۲۰۰۰، ترجمه شریفی و همکاران، ۱۳۸۸)؛ که از جمله این روش‌ها می‌توان به روش منتل هانسزل، رگرسیون لجستیک و دشواری سؤال تبدیل شده[2] اشاره کرد. در چارچوب رویکرد سؤال – پاسخ نیز چندین روش برای ارزیابی افت پارامترها مطرح شده است که این روش‌ها در دو گروه کلی مساحت میان منحنی ویژگی سؤال گروه مرجع و کانونی و روش آزمون‌های آماری یکسانی پارامترهای سؤال طبقه‌بندی می‌شوند (شولتز، ویتنی و زیکار[3]، ۲۰۱۳).

با توجه به مبانی نظری و تجربی و اهمیتی که افت پارامترها در آزمون‌های تیمز دارد پژوهش حاضر نیز در صدد پاسخگویی به پرسش‌های زیر است:

۱- آیا پارامتر دشواری سؤال‌های ریاضیات تیمز پایه هشتم از سال ۲۰۰۳ تا ۲۰۱۱ افت پیداکرده است؟

۲- آیا پارامتر تشخیص سؤال‌های ریاضیات تیمز پایه هشتم از سال ۲۰۰۳ تا ۲۰۱۱ افت پیداکرده است؟

۳- آیا افت پارامترهای سؤال‌ها بر برآورد پارامترهای توانایی دانش‌آموزان در درس ریاضی تیمز پایه هشتم تأثیر می‌گذارد؟

روش پژوهش

جامعه آماری شامل همه شرکت‌کنندگان در آزمون ریاضیات پایه هشتم تیمز ایران بود که طی سال‌های ۲۰۰۳، ۲۰۰۷ و ۲۰۱۱ در این آزمون شرکت کرده بودند. برای افزایش دقت در نمونه‌گیری، پس از حذف مواردی که به هیچ‌یک از ٤۰ سؤال روند پاسخ صحیح نداده بودند، تعداد ۱۲۲٤ نفر (٤۰۸ نفر در هر دوره) به شیوه سرشماری

---

1. Lord & Novick
2. Transformed Item Difficulty
3. Shultz, Whitney & Zickar

به‌عنوان نمونه انتخاب شدند. با توجه به این روش از همه وزن‌های به کار برده شده در فرایند نمونه‌گیری آزمون تیمز در این پژوهش نیز استفاده شد.

برای ارزشیابی پیشرفت تحصیلی در مقیاس گسترده معمولاً مجموعه‌ای از سؤال‌ها که «لنگر»[1] و در آزمون تیمز «روند» نامیده می‌شود، در دوره‌های مختلف زمانی و اجراهای مختلف تکرار و به‌منظور همترازسازی به کار می‌روند، استفاده می‌شود. این سؤال‌ها مبنایی برای مقایسه پیشرفت تحصیلی کشورها با کشورهای دیگر و نیز پیشرفت خود نسبت به سال‌های گذشته می‌شود. در آزمون تیمز نیمی از سؤال‌ها هر دوره به‌عنوان سؤال‌های روند برای دوره بعدی، حفظ و بقیه منتشر می‌شوند؛ به‌طوری که این سؤال‌ها در سه دوره، حفظ و در سال سوم همه سؤال‌ها منتشر می‌شود و سؤال‌های جدید جایگزین قبلی‌ها می‌شود و این روند ادامه پیدا می‌کند؛ اما تعداد سؤال‌های روند در طول دوره‌های مختلف تغییر می‌کند.

به‌منظور سهولت در تحلیل داده‌ها و به پیروی از مینائی (۱۳۹۱) و لی، پارک و تایالان[2] (۲۰۱۱) سؤال‌های چندارزشی با کدگذاری مجدد[3] به سؤال‌های دوارزشی (۰ و ۱) تبدیل شد. به این منظور، اگر فردی در یک سؤال چندارزشی بالاترین نمره را کسب کرده بود نمره ۱ و چنانچه نمره کامل نگرفته بود نمره صفر به او تعلق گرفت. سؤال‌هایی که افراد آنها را بدون پاسخ[4] گذاشته بودند در چهارچوب تیمز آنها را سؤال‌های «بدون پاسخ» و نیز سؤال‌هایی که به دلیل کمبود زمان فرصت پاسخ‌دهی به آنها نبوده است، سؤال‌های «جامانده»[5] می‌نامند، در این مورد از روش دومرحله‌ای (لادلو و اولیری[6]، ۱۹۹۹) استفاده شد. در این راهبرد که در مطالعات تیمز نیز از آن استفاده می‌شود در مرحله اول که هدف آن برآورد پارامترهای سؤال است، سؤال‌های بدون پاسخ به‌صورت سؤال‌های نادرست (نمره صفر) و سؤال‌های جامانده به‌صورت سؤال‌های اجرانشده، کدگذاری می‌شوند و پارامترهای سؤال برآورد می‌شود؛ در مرحله دوم، پارامترهای سؤال که در مرحله اول به‌دست آمده‌اند به‌عنوان پارامترهای

---

1. Anchor
2. Lee, Park & Taylan
3. Recoding
4. Omitted
5. Not-reached
6. Ludlow & O'leary

ثابت و معلوم فرض می‌شوند و پارامترهای توانایی افراد برآورد می‌شود. در این مرحله که هدف آن برآورد توانایی افراد است، سؤال‌های بدون پاسخ و جامانده به‌عنوان سؤال‌های نادرست در نظر گرفته می‌شود و به آنها نمره صفر تعلق می‌گیرد.

برای تجزیه‌وتحلیل یافته‌ها، نخست مفروضه تک‌بعدی بودن[1] بررسی شد. از شیوه‌های بررسی مفروضه تک‌بعدی بودن، روش‌های مبتنی بر استقلال موضعی[2] است که از انواع آن می‌توان به تحلیل عاملی خطی[3] (همبلتون و تراب[4]، ۱۹۷۳؛ رکاس[5]، ۱۹۷۹)، تحلیل عاملی غیرخطی[6] (گساورلی و دی چمپلین[7]، ۱۹۹٦) و تحلیل عاملی با اطلاعات کامل[8] (مک‌لاود، سویگرت و تیسن[9]، ۲۰۰۱) اشاره کرد؛ اما مشکل اصلی در تحلیل عاملی خطی، خطی بودن آن است. با توجه به این موضوع و مشکلات ناشی از استفاده از ماتریس‌های همبستگی فی و تتراکوریک، همبلتون و راوینلی[10] (۱۹۸٦) پیشنهاد کرده‌اند که برای بررسی تک‌بعدی بودن داده‌های تستی از روش تحلیل عاملی غیرخطی که مک‌دونالد[11] (۱۹٦۷،۱۹۸۲) ارائه کرده است، استفاده شود. این رویکرد با نرم‌افزار NOHARM (فریزر[12]، ۱۹۸۸) قابل اجراست (مینائی و فلسفی‌نژاد، ۱۳۸۹). با توجه به پیشنهادهای مینائی و فلسفی‌نژاد (۱۳۸۹) و برتری‌های این روش با استفاده از نرم‌افزار NOHARM مفروضه تک‌بعدی بودن بررسی شد.

با توجه به مزیت‌هایی که روش آماری رگرسیون لجستیک دارد روش مناسبی برای بررسی افت پارامترها به شمار می‌رود که در این پژوهش با استفاده از نرم‌افزار

---

1. Unidimensionality
2. Local Independence
3. Linear Factor Analysis
4. Hambleton & Traub
5. Reckase
6. Non-Linear
7. Gessaroli & De Champlain
8. Full Information Factor Analysis
9. McLeod, Swygert & Thissen
10. Hambleton & Rovinelli
11. McDonald
12. Fraser

آماری SPSS۲۳ و دستور نوشته شده (macro) زومبو[1] (۱۹۹۹) اجرا شد. بر اساس وو و همکاران[2] (۲۰۰۶) نخستین مزیت روش رگرسیون لجستیک مانند رگرسیون چندمتغیری استفاده از متغیر گروهی به تعداد دلبخواه است؛ یعنی برخلاف بررسی کارکرد افتراقی در گروه‌هایی مانند جنسیت با دو گروه می‌توان در این روش بیش از دو گروه را در معادله وارد کرد؛ دومین مزیت آن، بررسی هم‌زمان افت یکنواخت و غیریکنواخت است؛ سومین مزیت این روش، فرض‌آزمایی و بررسی میزان اندازه اثر است. روش رگرسیون لجستیک یک روش سلسله‌مراتبی سه‌مرحله‌ای برای ورود داده‌ها با در نظر گرفتن متغیر سؤال (item) به‌عنوان متغیر وابسته است که این مراحل به‌صورت زیر ارائه می‌شوند:

مرحله اول: ابتدا نمره‌ای که پاسخ‌دهنده در کل آزمون کسب کرده است به‌عنوان متغیر همتاسازی[3] وارد معادله شد؛

مرحله دوم: متغیر گروهی (در این پژوهش دوره‌های مختلف زمانی) وارد شد؛

مرحله سوم: تعامل میان دوره‌های زمانی و نمره کل وارد معادله شد.

معادله رگرسیون برای هرکدام از مراحل بالا به شکل زیر است:

Model(1): Logit = b0 + b1*Total (df = 1)
Model(2): Logit = b0 + b1*Total + b2*Cycle (df = 1+2= 3)
Model(3): Logit = b0 + b1*Total + b2*Cycle + b3*Total by Cycle (df = 1+2+2= 5)

که b_0 پارامتر عرض از مبدأ، b_1 ضریب رگرسیون برای متغیر همتاسازی آزمون، b_2 ضریب رگرسیون برای نشانگر عضویت گروهی (سه دوره زمانی) و b_3 ضریب رگرسیون برای تعامل میان عضویت گروهی و متغیر همتاسازی است. درجات آزادی در این طرح در مرحله اول با (df = 1) و با اضافه شدن ۳ دوره زمانی در مرحله دوم (df = 1+2= 3) و در مرحله آخر در تعامل میان آنها (df = 1+2+2= 5) حاصل می‌شود (وو و همکاران، ۲۰۰۶).

---

1. Zumbo
2. Wu et al
3. Matching variable

با استفاده از مقادیر χ^2 محاسبه شده برای هر مرحله می‌توان معنی‌داری افت پارامترها و نوع آن را مشخص کرد؛ به‌این‌ترتیب که با کم کردن مقدار χ^2 مرحله سوم از مرحله اول با ٤ درجه آزادی (کسر مرحله سوم، ۵ درجه آزادی و مرحله اول، ۱ درجه آزادی) می‌توان در صورت معنی‌داری وجود افت را تعیین و نیز برای آزمودن نوع افت (یکنواخت و غیریکنواخت بودن) از تفاوت میان مرحله سوم و دوم می‌توان فهمید چه مقدار از افت غیریکنواخت است. در صورت معنی‌داری با افت غیریکنواخت و در غیر اینصورت با افت یکنواخت روبرو هستیم (سوامیناتان و راجرز[1]، ۱۹۹۰)؛ روش دیگر از تفاضل R^2 مرحله دوم از مرحله اول برای شناسایی افت یکنواخت است؛ به این صورت که اگر تفاضل آن دو معنی‌دار باشد افت از نوع یکنواخت است. برای بهبود بخشیدن به روش آماری با استفاده از کسر نمره کل از نمره فرد در آزمون، نمره‌های کل خالص[2] شدند. با توجه به موارد یاد شده، مقدار اندازه اثر افت پارامتر نیز از کسر مقدار R^2 مرحله سوم از مرحله اول به دست می‌آید و نیز برای به دست آوردن مقدار بزرگی نوع افت (یکنواخت و غیریکنواخت) R^2 مرحله سوم را از مرحله دوم یا مرحله دوم از اول کسر شد تا مقدار بزرگی غیریکنواخت بودن یا یکنواخت بودن آن حاصل شود. اگر تفاوت گروه مرجع با گروه کانونی در مقدار پارامتر دشواری سؤال باشد، سؤال دارای افت یکنواخت است. درصورتی‌که اختلاف در پارامتر قدرت تشخیص باشد، افت پارامتر غیریکنواخت را وجود می‌آورد، با توجه به حجم بالای نمونه، اندازه اثرهای کوچک نیز از لحاظ آماری معنی‌دار شدند.

پس از شناسایی سؤال‌های دارای افت، برای بررسی تأثیر این میزان افت روی برآورد توانایی دانش‌آموزان، ابتدا توانایی (θ) دانش‌آموزان با استفاده از نرم‌افزار بایلوگ (BILOG-MG) برای کل سؤال‌ها محاسبه شد؛ سپس سؤال‌های دارای افت، حذف و دوباره توانایی دانش‌آموزان ثبت شد. برای بررسی معنی‌داری تفاوت میانگین توانایی‌ها از آزمون t وابسته استفاده شد؛ در صورت معنی‌داری افت پارامترها روی برآورد توانایی دانش‌آموزان تأثیر داشته است.

---

1. Swaminathan & Rogers
2. purified total

یافته‌ها

از مفروضه‌های مهم برای بررسی افت پارامترهای سؤال تک‌بعدی بودن آزمون است. نرم‌افزار NOHARM ریشه دوم میانگین مجذورات پس‌مانده‌های RMSR را محاسبه می‌کند و به‌عنوان شاخصی برای برازش مدل ارائه می‌دهد.

درواقع، RMSR با ریشه دوم میانگین مجذورات تفاوت بین کوواریانس‌های مشاهده شده و کوواریانس‌های پیش‌بینی شده برابر است. پس می‌توان گفت، مقادیر کوچک RMSR نشان‌دهنده برازش مدل با داده‌ها است. شاخص دیگر برای بررسی برازش مدل، شاخص خوبی برازندگی GFI تاناکا[1] (۱۹۹۳) است. مک دونالد (۱۹۹۷) پیشنهاد کرده مقدار ۰/۹۰ برای این شاخص نشان‌دهنده برازش قابل‌قبول و مقدار ۰/۹۵ بیانگر برازش خوب مدل با داده‌ها است. اگر GFI=1 باشد بیانگر برازش کامل است (مینائی و فلسفی‌نژاد، ۱۳۸۹).

جدول (۲) شاخص‌های برازش تک‌بعدی بودن دفترچه‌های آزمون در سه دوره

GFI	RMSR	SRS	Cycle
۰/۹۰٤	۰/۰۱	۰/۰۸٦	۲۰۰۳
۰/۸۷	۰/۰۱	۰/۱	۲۰۰۷
۰/۸۸	۰/۰۱	۰/۱	۲۰۱۱

ملاک‌های مک‌دونالد (۱۹۹۷) برای مطلوبیت برازش مدل با داده‌ها مقادیر نسبی هستند و ملاک مطلق در این زمینه وجود ندارد، اما با این وجود مقادیر به دست آمده تفاوت ناچیزی با ملاک‌های ارائه شده دارند (۰/۰۲ و ۰/۰۳)؛ بنابراین همان‌طور که نتایج جدول (۲) نشان می‌دهد هر سه دوره برگزاری آزمون شاخص‌های برازش نسبتاً مطلوبی دارند و شواهدی برای رد مفروضه تک‌بعدی بودن سؤال‌های روند آزمون ریاضیات تیمز پایه هشتم طی سه دوره برگزاری از سال ۲۰۰۳ تا ۲۰۱۱ وجود نداشت. پس از رعایت مفروضه تک‌بعدی بودن، برای پاسخگویی به سؤال‌های پژوهش از روش آماری رگرسیون لجستیک استفاده شد. به نظر زومبو (۱۹۹۹) طبق

[1]. Tanaka

این روش برای اینکه سؤال ازنظر آماری دارای افت شناخته شود باید تفاوت مرحله سوم از اول (P≤۰/۰۰۱) داشته باشد؛ همچنین برای بررسی میزان بزرگی افت در پارامترهای سؤال از ملاک اندازه اثر جودوین و جیرل[1] (۲۰۰۱) استفاده شد؛ به این صورت که هرگاه تغییرات R^2 ناجلکرک[2] بین مرحله دوم و اول (برای بررسی افت یکنواخت) و همچنین مرحله سوم از دوم (برای بررسی افت غیریکنواخت) کمتر از ۰/۳۵ باشد مقدار آن ناچیز و مقادیر بین ۰/۳۵ تا ۰/۷ به‌عنوان متوسط و بزرگ‌تر از ۰/۷ مقدار بزرگ محسوب می‌شود. بر این اساس، دو سؤال از مجموعه سؤال‌های روند آزمون تیمز در سه دوره نامبرده دارای افت یکنواخت شناخته شدند که هر دو سؤال اندازه اثر ناچیزی داشتند. نتایج این تحلیل‌ها در جدول (۳) ارائه شده است.

جدول (۳) نتایج رگرسیون لجستیک در خصوص وضعیت افت در سؤال‌های روند تیمز ۲۰۰۳ تا ۲۰۱۱

شماره سؤال	گام اول		گام دوم		گام سوم		گام سوم-گام اول			گام دوم-گام اول		
	χ^2	R^2	χ^2	R^2	χ^2	R^2	χ^2	R^2	P	χ^2	R^2	P
M032166	٤٥/٦	۰/۰٥	٤٥/۷	۰/۰٥	٤٨	۰/۰٥	۲/٤	۰/۰	۰/۷	۰/۱	۰/۰	
M032721	۲٦/۲	۰/۰۳	۳۰/۲	۰/۰٤	۳٤/٤	۰/۰٤	۸/۲	۰/۰۱	۰/۰۸	٤	۰/۰۱	
M032757	٦۲/۱	۰/۰۷	۸۲	۰/۰۹	۸٤	۰/۰۹	۲۱/۹	۰/۰۲	۰/۰۰۱	۲۰	۰/۰۲	۰/۰۰۱
M032760a	۱۰۷/٤	۰/۱٥	۱۲۰	۰/۱۷	۱۲۰	۰/۱۷	۱۲/٦	۰/۰۲	۰/۰۱۳	۱۲/٦	۰/۰۲	
M032760b	۸۰/۳	۰/۱۳	۸۸	۰/۱٥	۹۰/٦	۰/۱٥	۱۰/۳	۰/۰۲	۰/۰۳٦	۷/۷	۰/۰۲	
M032760c	٦۲/۱	۰/۱٥	۸٥	۰/۲۱	۸۷/۷	۰/۲۲	۲٥/٦	۰/۰۷	۰/۰۰۱	۲۲/۹	۰/۰٦	۰/۰۰۱
M032761	٥۰/۱	۰/۲٥	٥٤/٥	۰/۲۷	٥٤/٦	۰/۲۷	٤/٥	۰/۰۲	۰/۳٤	٤/٤	۰/۰۲	
M032692	۹۰/۸	۰/۱۲	۹٥/۳	۰/۱۳	۹٥/۷	۰/۱۳	٤/۹	۰/۰۱	۰/۳	٤/٥	۰/۰۱	
M032626	۷٥	۰/۰۸	۸٤/۲۷	۰/۰۹	۸۸	۰/۰۱	۱۳	۰/۰۲	۰/۰۱	۹	۰/۰۱	
M032595	۸٦/٤	۰/۱	۹٥/۳	۰/۱	۹۷/۳	۰/۱۱	۱۰/۹	۰/۰۱	۰/۰۳	۸/۹	۰/۰	
M032673	٦۷/۷	۰/۰۸	۷۰	۰/۰۸	۷۱/۲	۰/۰۸	۳/٥	۰/۰	۰/٤۸	۲/۳	۰/۰	
M032094	٤۱/٥	۰/۰٥	٤۲/۳	۰/۰٥	٤٥/۳	۰/۰٥	۳/۸	۰/۰۳	۰/٤۹	۰/۸	۰/۰	
M032662	٥	۰/۰۱	۱٤/۲	۰/۰۲	۱۹	۰/۰۳	۱٤	۰/۰۲	۰/۰۰۷	۹/۲	۰/۰۱	

[1]. Jodoin & Gierl
[2]. Nagelkerke

شماره سؤال	گام اول		گام دوم		گام سوم		گام سوم-گام اول			گام دوم-گام اول		
	χ^2	R^2	χ^2	R^2	χ^2	R^2	χ^2	R^2	P	χ^2	R^2	P
M032064	۱۳۱	۰/۱۹	۱۳۲	۰/۱۹	۱۳۵	۰/۱۹	٤	۰/۰	۰/٤	۱	۰/۰	
M032419	۱٦/۳	۰/۰۲	۱٦/۸	۰/۰۲	۲۲/٤	۰/۰۳	٦/۱	۰/۰۱	۰/۲	۰/۵	۰/۰	
M032477	۵۱/۷	۰/۰٦	۵٤/۳	۰/۰٦	۵۹/۷	۰/۰۷	۸	۰/۰۱	۰/۰۹	۲/٦	۰/۰	
M032538	۰/۱٦	۰/۰	۲/۳	۰/۰	۸/۵	۰/۰	۸/۳٤	۰/۰	۰/۰۸	۲/۱٤	۰/۰	
M032324	۲٤/۲	۰/۰۳	۲۸/۲	۰/۰۳	۲۸/۳	۰/۰۳	٤	۰/۰	۰/٤	٤	۰/۰	
M032116	۵۲/۵	۰/۰٦	۵۲/۹	۰/۰٦	۵۳	۰/۰٦	۰/۵	۰/۰	۰/۹۷	۰/٤	۰/۰	
M032100	٦۰	۰/۰۷	٦۱/۵	۰/۰۷	٦۵/٤	۰/۰۸	۵/٤	۰/۰۱	۰/۲۵	۱/۵	۰/۰	
M032403	۲٦	۰/۰۳	۲۷/۵	۰/۰۳	۲۷/۹	۰/۰۳	۱/۹	۰/۰	۰/۷	۱/۵	۰/۰	
M032734	٤۹/۵	۰/۰۵	۵۲/٦	۰/۰٦	۵۵/۲	۰/۰٦	۵/۷	۰/۰۱	۰/۲۳	۳/۱	۰/۰۱	
M032397	۱۲	۰/۰۱	۱۲/۱	۰/۰۱	۱۸/۳	۰/۰۲	٦/۳	۰/۰۱	۰/۱۸	۰/۱	۰/۰	
M032695	۱۲۷	۰/۱٤	۱۳٤	۰/۱۵	۱۳۹	۰/۱۵	۱۲	۰/۰۱	۰/۰۲	۷	۰/۰۱	
M032132	۳٤/۹	۰/۰٤	۳٦/٤	۰/۰٤	۳۸/٦	۰/۰٤	۳/۷	۰/۰	۰/٤	۱/۵	۰/۰	
M032352	۱٦/۲	۰/۰۲	۲۲/۲	۰/۰۳	۲۳/۷	۰/۰۳	۷/۵	۰/۰۱	۰/۱۱	٦	۰/۰۱	
M032735	۸۰/۱	۰/۲۳	۸۱/٦	۰/۲۳	۸۵/٤	۰/۲۵	۵/۳	۰/۰۲	۰/۲٦	۱/۵	۰/۰	
M032683	۱۱۳/۷	۰/۱۹	۱۱٤	۰/۱۹	۱۱۵	۰/۱۹	۱/۳	۰/۰	۰/۸٦	۰/۳	۰/۰	
M032738	۸۲	۰/۰۹	۸۳	۰/۰۹	۸۳/۸	۰/۰۹	۱/۸	۰/۰	۰/۷۷	۱	۰/۰	
M032295	۱۱۸	۰/۱۲	۱۲٦	۰/۱۳	۱۲٦	۰/۱۳	۸	۰/۰۱	۰/۰۹	۸	۰/۰	
M032331	۸/۳	۰/۰۱	۹/۱	۰/۰۱	۹/۵	۰/۰۱	۱/۲	۰/۰	۰/۸۷	۰/۸	۰/۰	
M032623	٦۲	۰/۰۸	٦۲/۹	۰/۰۸	٦۳/۱	۰/۰۸	۱/۱	۰/۰	۰/۸۹	۰/۹	۰/۰	
M032697	۸۱	۰/۰۹	۸۳/۵	۰/۰۹	۸٤/۹	۰/۰۹	۳/۹	۰/۰	۰/٤۲	۲/۵	۰/۰	
M032047	۸/۱	۰/۰۱	۸/۱	۰/۰۱	۸/۷	۰/۰۱	۰/٦	۰/۰	۰/۹٦	۰/۰	۰/۰	
M032398	۲۹/۱	۰/۰۳	۳۵/۳	۰/۰٤	۳٦/۸	۰/۰٤	۷/۷	۰/۰۱	۰,۱	٦/۲	۰/۰۱	
M032507	۱۰/۸	۰/۰۱	۱۱/۷	۰/۰۲	۱۱/۹	۰/۰۲	۱/۱	۰/۰۱	۰/۹	۰/۹	۰/۰۱	
M032424	٤٦	۰/۰۵	٤۹/۵	۰/۰٦	۵۰/۷	۰/۰٦	٤/۷	۰/۰۱	۰/۳۲	۳/۵	۰/۰۱	
M032681a	٤۳/۵	۰/۰۵	٤٦	۰/۰۵	٤۹	۰/۰۵	۵/۵	۰/۰	۰/۲٤	۲/۵	۰/۰	
M032681b	۳٦	۰/۰٤	۳۷/۸	۰/۰۵	۳۸	۰/۰۵	۲	۰/۰۱	۰/۷۵	۱/۸	۰/۰۱	
M032681c	۳۵/۲	۰/۰٦	۳۸	۰/۰۷	۳۸	۰/۰۷	۲/۸	۰/۰۱	۰/٦	۲/۸	۰/۰۱	

پرسش اول: آیا پارامتر دشواری سؤال‌های ریاضیات تیمز پایه هشتم از سال ۲۰۰۳ تا ۲۰۱۱ افت پیداکرده است؟

همان‌طور که گفته شد، هنگامی‌که افت پارامترها از نوع یکنواخت باشد به این معنی که شیب نمودارها در دوره‌های مختلف تفاوت چشم‌گیری نداشته باشد و از نظر آماری، تفاوت مرحله دوم از اول معنی‌دار و چشم‌گیر باشد تغییرات تنها در پارامتر دشواری رخ داده است. بررسی‌ها نشان داد دو سؤال از سه سؤال که به‌عنوان سؤال‌های دارای افت شناخته شد، از نوع یکنواخت هستند. به این معنی که پارامترهای دشواری طی سه دوره افت پیدا کرده است، اما این سؤال‌ها اندازه اثر ناچیزی داشتند. در جدول (٤) درجه دشواری سؤال‌هایی که افت را در خود نشان داده‌اند بر اساس میانگین جهانی و کشور ایران بر اساس نظریه کلاسیک بیان شده است.

جدول (٤) درجه دشواری سؤال‌های دارای افت به تفکیک سال برگزاری

	۲۰۰۳ ایران	۲۰۰۳ بین‌الملل	۲۰۰۷ ایران	۲۰۰۷ بین‌الملل	۲۰۱۱ ایران	۲۰۱۱ بین‌الملل
M032757	۰/۳۵	۰/۵	۰/٤۹	۰/۵۲	۰/۵۱	۰/۵٤
M032760c	۰/۰۷	۰/۱۲	۰/۰٦	۰/۱۲	۰/۰۷	۰/۱۵
M032662	۸/۷	۲۰/۵	۱۵/۹	۲٤/۲	۱۵/۸	۲٤/۳

پرسش دوم: آیا پارامتر تشخیص سؤال‌های ریاضیات تیمز پایه هشتم از سال ۲۰۰۳ تا ۲۰۱۱ افت پیدا کرده است؟

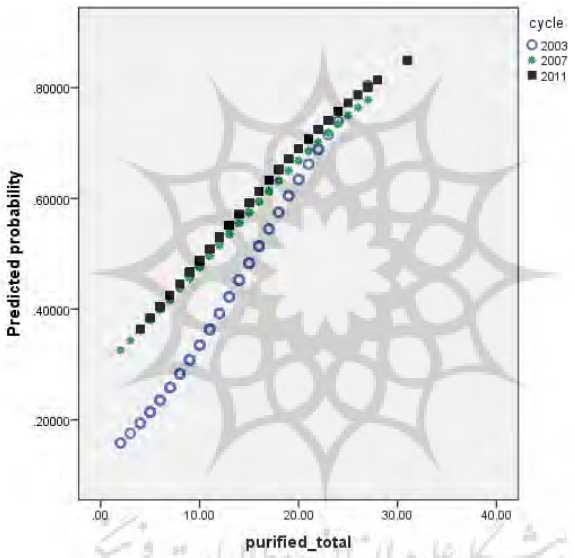
همان‌طور که تجزیه‌وتحلیل داده‌های پژوهش نشان داد هیچ‌یک از سؤال‌های پژوهش ازنظر قدرت تشخیص افت را در خود نشان نداده بودند؛ به‌عبارت دیگر، افت غیریکنواخت در سؤال‌های آزمون شناسایی نشد.

پرسش سوم: آیا افت پارامترهای سؤال‌ها بر برآورد پارامترهای توانایی دانش آموزان در درس ریاضی تیمز پایه هشتم تأثیر می‌گذارد؟

برای پاسخ به این پرسش، نخست پارامتر توانایی (θ) دانش‌آموزان با استفاده از نرم‌افزار BILOG-MG در حالتی که تمام سؤال‌های روند وجود داشت محاسبه شد؛ سپس، سؤال‌های دارای افت از مجموعه سؤال‌ها حذف و دوباره پارامتر توانایی (θ) هر دانش‌آموز در غیاب این سؤال‌ها حساب شد. برای بررسی میزان تغییرات و معنی داری میانگین توانایی‌ها در هر دو حالت با استفاده از روش آماری t وابسته

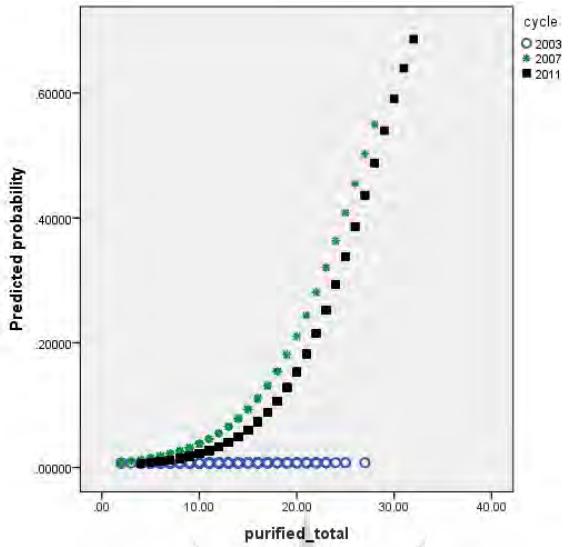
تجزیه‌وتحلیل شد. نتیجه آنکه با درجه آزادی ۱۲۲۲ و ۹۵ درصد اطمینان بین میانگین توانایی دانش‌آموزان، قبل و بعد از حذف سؤال‌های دارای افت با (t=۰/٤۹۳) ازنظر آماری (P=۰/٦۲) تفاوت معنی‌داری دیده نشد. با توجه به معنی‌دار نبودن این آزمون می‌توان نتیجه گرفت سؤال‌های دارای افت در این آزمون در برآورد توانایی دانش آموزان تأثیر نداشته است.

پس از پاسخ به پرسش‌های پژوهش و بررسی همه سؤال‌های روند در ادامه، سؤال‌هایی که الگویی از افت را در خود نشان داده‌اند، بررسی می‌شوند. سؤال ۰۳۲۷۵۷ M با (P≤۰/۰۰۱) و اندازه اثر ($\mathcal{R}^2$=۰/۰۲) دارای افت پارامتر یکنواخت شناخته شد که طبق ملاک اندازه اثر جودوین و جیرل دارای مقادیر ناچیزی هستند.



نمودار (۱) پراکندگی نمره‌های سه دوره مربوط به سؤال ۰۳۲۷۵۷ M

سؤال ۰۳۲۷٦۰c M با (P≤۰/۰۰۱) و اندازه اثر ($\mathcal{R}^2$=۰/۰۷) دارای افت پارامتر یکنواخت شناخته شد که طبق ملاک اندازه اثر جودوین و جیرل دارای مقادیر ناچیزی هستند.



نمودار (۲) پراکندگی نمره‌های سه دوره مربوط به سؤال M ۰۲۳۷۶۰c

بحث و نتیجه‌گیری

لین و گرانلند (۲۰۰۰) روایی را «یک ارزشیابی از کفایت و مناسبت تفسیرها و استفاده‌های نتایج سنجش» تعریف کرده‌اند (نقل از سیف، ۱۳۹۰). ازجمله مواردی که خطر جدی برای روایی آزمون به شمار می‌رود، افت در پارامترهای سؤال است. در همین خصوص، در پژوهش حاضر نیز وجود افت در پارامتر سؤال‌های روند تیمز در طی سه دوره ۲۰۰۳، ۲۰۰۷ و ۲۰۱۱ بررسی شد.

نتایج پژوهش نشان داد که از بین ٤٠ سؤال روند آزمون ریاضیات پایه هشتم دو سؤال، دارای افت پارامتر بودند. هر دو سؤال (M ۰۲۳۷٦۰c و M ۰۳۲۷۵۷) از نوع یکنواخت بودند اما با توجه به ملاک اندازه اثر جودوین و جیرل هر دو سؤال اندازه اثر ناچیزی داشتند. هدف و انگیزه دانش‌آموزان از یادگیری مفاهیم درسی در دوره‌های مختلف، با توجه به تغییراتی که در نظام آموزشی هر کشوری به وجود می‌آید، تغییر می‌کند. در ایران نیز با توجه به سرنوشت‌ساز بودن و اهمیتی که قبولی در آزمون کنکور وجود دارد، در شیوه یادگیری مفاهیم درسی دانش‌آموزان تغییراتی به وجود آمده است، به‌طوری که در ساعات غیردرسی و در کلاس‌های آمادگی کنکور بیشترین تأکیدی که بر مفاهیم می‌شود در جهت کسب نتیجه‌ای بهتر است و فراگیری مطالب از

دانش‌محور بودن به مهارت‌محور بودن تبدیل شده است. به همین سبب می‌توان ازجمله مهم‌ترین علل در افت پارامترهای سؤال‌ها را نیز به همین موضوع نسبت داد.

افت در پارامترهای سؤال به دلایل مختلف رخ می‌دهد، اما در عمل احتمالاً به علت استفاده بیش‌ازحد آزمون در طول زمان و تغییرات برنامه درسی اتفاق می‌افتد (باک و همکاران، ۱۹۸۸). همچنین این پدیده ممکن است به علت امنیت و کنترل ضعیف در نگهداری سؤال‌ها و فاش شدن آزمون اتفاق بیفتد. یکی دیگر از منابع انحراف می‌تواند تغییر برنامه درسی باشد (گلدشتاین، ۱۹۸۳). در شناسایی علل افت پارامترهای سؤال باید به دنبال پاسخگویی به پرسش‌های زیر بود:

آیا تأکیدی که بر محتوای سؤال در حال حاضر می‌شود، در مقایسه با زمانی که سؤال طراحی شده، تغییر کرده است؟

آیا ممکن است شرایط انگیزشی برای پاسخ‌دهندگان در زمان‌های مختلف تغییر کرده باشد؟

آیا تغییر یا بازبینی مجددی در متن اصلی یا گزینه‌ها در بانک سؤال‌ها صورت گرفته است؟

آیا محتوای سؤال ازنظر درسی منسوخ‌شده است؟

آیا تغییری در مدت‌زمان پاسخگویی به سؤال‌ها به وجود آمده است؟

اگر هرکدام از شرایط ذکر شده وجود داشته باشد، سؤال باید به‌عنوان یک سؤال جدید در نظر گرفته شود که دارای پارامترهایی متفاوت از سؤال اصلی است و باید پارامترهای جدیدی برای این سؤال، برآورد و جایگزین سؤال قبلی شود تا در آینده به‌عنوان سؤالی با پارامترهای جدید استفاده شود؛ اما در عمل کارشناسان آزمون توصیه به حذف یا برآورد دوباره سؤال‌های دارای افت می‌کنند (کولن و برنان، ۱۹۹۵).

بنا به دلایلی که ذکر شد اغلب سؤال‌های آزمون به‌ناچار به‌مرور زمان منسوخ می‌شوند و به تجدیدنظر و کنار گذاشتن از خزانه سؤال نیاز دارند. به‌روز کردن مکرر سؤال‌های آزمون، امری مطلوب و پسندیده است، به این دلیل که محتوای سؤال‌های آزمون را از تغییرات برنامه درسی و تغییرات در نظام آموزش‌وپرورش و استفاده بیش‌ازحد سؤال در آزمون حفظ می‌کند. این فرایند شامل کنارگذاری یا جانشینی سؤال‌های انتخاب‌شده در یک آزمون یا در یک مجموعه آزمون است؛ اما پیش از کنار گذاشتن سؤال‌ها باید از سودمند نبودن آنها اطمینان حاصل کرد. به گفته گلدشتاین

(۱۹۸۳)، به‌منظور ارزیابی درست از سودمند بودن سؤال‌های آزمون، پژوهشگران باید چند پرسش زیر را مدنظر قرار دهند.

-آزمون‌گیرنده تا چه اندازه می‌تواند یک سؤال را در آزمون‌های بعدی تکرار کند تا آن سؤال منسوخ شود؟

-چه زمانی یک آزمون باید مورد ارزیابی مجدد قرار گیرد؟

-چه هنگام پژوهشگر می‌تواند ادعایی معتبر در خصوص تغییرات پاسخ آزمودنی‌ها در طول زمان داشته باشد؟

در این رابطه، در پژوهش‌هایی با روش‌های مختلف، افت پارامترهای سؤال بررسی شد؛ برای مثال، در زمینه افت پارامترهای سؤال در آزمون تیمز، وو و همکاران (۲۰۰٦) افت پارامترهای سؤال در آزمون ریاضیات تیمز و مقایسه کشور سنگاپور و آمریکا را بررسی کردند. ایشان سه دوره آزمون تیمز را در طی سال‌های ۱۹۹۵ تا ۲۰۰۳ با استفاده از روش رگرسیون لجستیک مورد بررسی قرار دادند. با توجه روش نمونه‌گیری و پاسخ دانش‌آموزان به‌صورت تصادفی به دفترچه‌های آزمون، نمره کل از حاصل جمع ۲۳ سؤال روند برای متغیر همتاسازی به دست آمد. سرانجام به این نتیجه رسیدند که با استفاده از ملاک ($P\leq$۰/۰۰۲) در هردو کشور سنگاپور و آمریکا سؤال‌های تیمز دارای افت پارامتر از نوع یکنواخت (۲۰ سؤال از ۲۳ سؤال برای سنگاپور و ۱۸ سؤال از ۲۳ سؤال برای آمریکا) و غیریکنواخت (۸ سؤال باز، ۲۳ سؤال برای سنگاپور و ۲۰ سؤال از ۲۳ سؤال برای آمریکا) بودند؛ اما با توجه به ناچیز بودن اندازه اثر طبق ملاک اندازه اثر جودوین و جیرل، مقدار افت برای تمامی سؤال‌ها ناچیز ($\mathcal{R}^2\leq$۰/۰۳۵) بود؛ درنتیجه سؤال‌های آزمون تیمز در این پژوهش بدون افت در نظر گرفته شد؛ اما باید دانست سؤال‌های روند در آزمون ذکر شده، ۲۳ سؤال بود اما در پژوهش حاضر ٤۰ سؤال روند وجود دارد. همچنین در آزمون ریاضیات پایه هشتم سال ۲۰۱۱ کشور سنگاپور رتبه ۲ و ایران رتبه ۳۲ از ٤۲ کشور را کسب کرده است و مقایسه این گروه‌ها با یکدیگر چندان مناسب نیست.

در مطالعه‌ای دیگر، پارک و همکاران (۲۰۱٦)، داده‌های حاصل از آزمون تیمز را طی سال‌های ۱۹۹۹ تا ۲۰۰۷ با استفاده از مدل‌های آمیخته[1] IRT بررسی کردند. نتایج

[1]. Mixture Models

بررسی‌های آنها تغییرات قابل توجهی در پارامترهای توانایی را در سه دوره برگزاری آزمون از خود نشان داد. وی[1] (۲۰۱۳) نیز تأثیر افت پارامترهای سؤال در برآورد توانایی دانش‌آموزان را در آزمون‌های چندمرحله‌ای بررسی کرد. وی سؤال‌ها را یک‌بار بدون وجود افت و بار دیگر با واردکردن درجات مختلف افت در سؤال‌ها بررسی کرد. نتایج این بررسی نشان داد افت سؤال باوجود تأثیر کمی که در درجات متفاوت روی برآورد توانایی می‌گذارد اما در کل افت پارامترها تأثیر معنی‌داری بر توانایی دانش‌آموزان ندارد.

همسو با نتایج این پژوهش، فن (۱۹۹۸)، مک‌دونالد و پائونونن (۲۰۰۲) و آددویین[2] (۲۰۱۰) تغییرناپذیری پارامتر تشخیص را ناچیز گزارش کردند و بیشتر تغییرات را در پارامتر دشواری گزارش کردند. در مطالعه‌ای، باک و همکاران (۱۹۸۸) افت پارامتر سؤال در آزمون بین‌المللی فیزیک که در طول ۱۰ سال اجرا شده بود را بررسی کردند و هیچ شاهدی بر وجود افت پارامتر قدرت تشخیص پیدا نکردند. از جمله محدودیت‌های این پژوهش، نبود پیشینه تجربی در زمینه افت پارامترهای سؤالات تیمز در کشور ایران بود که در نتیجه امکان مقایسه نتایج فراهم نشد؛ همچنین پیشنهاد می‌شود در پژوهش‌های بعدی از روش‌های مبتنی بر نظریه‌های جدید اندازه‌گیری نیز استفاده شود تا امکان مقایسه توان روش‌های آماری برقرار شود.

1. Wei
2. Adedoyin

منابع

امبرتسون، سوزان ای. و رایس، استیون پی. (۲۰۰۰). *نظریه‌های جدید روان‌سنجی برای روان شناسان (IRT)*؛ ترجمه حسن پاشا شریفی، ولی‌الله فرزاد، مجتبی حبیبی و بلال ایزانلو (۱۳۸۸). تهران: انتشارات رشد.

سیف، علی‌اکبر (۱۳۹۰). *اندازه‌گیری، سنجش و ارزشیابی آموزشی*. تهران: انتشارات دوران، ویرایش ششم.

میرزاخانی، علیرضا و فرزاد، ولی‌الله (۱۳۹۲). ساخت مقیاس حل مسئله تیمز ۲۰۰۷ و بررسی موفقیت در حل مسئله دانش‌آموزان پایه سوم راهنمایی ایران در تیمز ۲۰۰۷. *فصلنامه تعلیم و تربیت*، ۲۹ (۲)، ۱۴۵-۱۶۴.

کبیری، مسعود (۱۳۹۲). کاربرد سنجش شناختی تشخیصی به‌منظور تعیین مهارت‌های کسب‌شده علوم تجربی در دانش‌آموزان پایه تحصیلی هشتم ایران بر اساس داده‌های تیمز ۲۰۱۱. *پایان‌نامه دکتری*، دانشگاه تهران.

کیامنش، علیرضا؛ صفرخانی، مریم؛ کبیری، مسعود و اقدسی، سمانه (۱۳۹۲). روند عملکرد علوم دانش آموزان سوم راهنمایی از ۱۳۷۸ تا ۱۳۸۶ با توجه به هدف‌های سند چشم انداز ۲۰ ساله: یافته های TIMSS در ایران و کشورهای منطقه. *مطالعات برنامه درسی*، ۳۱، ۶۹ – ۹۰.

مینائی، اصغر (۱۳۹۲). سنجش مقایسه پذیری سازه و تحلیل کارکرد افتراقی سؤال‌ها (DIF) و بلوک‌های (DTF) آزمون علوم پایه هشتم تیمز ۲۰۰۷ در بین دانش آموزان ایران و آمریکا. *فصلنامه اندازه‌گیری تربیتی*، ٤ (۱۱)، ۱۰۹-۱۴۶.

مینائی، اصغر (۱۳۹۱). *مدل‌پردازی تشخیصی شناختی (CDM) سؤال‌های ریاضیات تیمز ۲۰۰۷ در دانش‌آموزان پایه هشتم ایران با استفاده از مدل یکپارچه با پارامترپردازی مجدد (RUM) و مقایسه مهارت‌های ریاضی دانش‌آموزان دختر و پسر*. رساله دکتری، دانشگاه علامه طباطبائی.

مینائی، اصغر و فلسفی‌نژاد، محمدرضا (۱۳۸۹). روش‌های سنجش تک‌بعدی بودن سؤال‌ها در مدل‌های دو ارزشی IRT. *فصلنامه اندازه‌گیری تربیتی*، ۳، ۷۹-۹۸.

Adedoyin, O. O. (2010). Investigating the invariance of person parameter estimates based on classical test and item response theories. *International Journal of Educational Sciences, 2* (2), 107-113.

Bock, R. D.; Murake, E. & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25* (4), 275-285.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement, 22* (1), 33-51.

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research, 29* (6), 543-553.

Fraser, C. (1988). *NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory. LKJ* [Computer software]. Armidale, Australia: The University of New England.

Gessaroli, M. E., & Champlain, A. F. (1996). Using an Approximate Chi-Square Statistic to Test the Number of Dimensions Underlying the Responses to a Set of Items. *Journal of Educational Measurement, 33* (2), 157-179.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement, 20* (4), 369-377.

Guo, F., & Wang, L. (2003, April). *Online calibration and scale stability of a CAT program*. In annual meeting of the National Council on Measurement in Education, Chicago: IL.

Hambleton, R. K. & Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10* (3), 287-302.

Hambleton, R. K. & Traub, R. E. (1973). Analysis of empirical data using two logistic latent trait models. *British Journal of Mathematical & Statistical Psychology,* 24, 273-281.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14* (4), 329-349.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, linking, and scaling: Methods and practices*. Springer-Verlag, New York.

Lee, Y. S.; Park, Y. S. & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the US national sample using the TIMSS 2007. *International Journal of Testing, 11* (2), 144-177.

Lord, F. M. & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA : Addison -Wesley.

Ludlow, L. H., & O'leary, M. (1999). Scoring omitted and not-reached items: Practical data analysis implications. *Educational & Psychological Measurement, 59* (4), 615-630.

McCoy, K. M. (2009). *The impact of item parameter drift on examinee ability measures in a computer adaptive environment.* (Unpublished doctoral dissertation). University of Illinois at Chicago, Chicago, IL.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric Monographs*, *15*, 32.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement*, *6,* 379-396.

McDonald, R. P. (1997). Normal-ogive multidimensional model. *Handbook of modern item response theory*, 257-269.

McLeod, L. D., Swygert, K. A. & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen & H. Wainer. (Ed.), *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum.

Mullis, I. V.; Martin, M. O.; Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics. International Association for the Evaluation of Educational Achievement*. Herengracht 487, Amsterdam, 1017 BT, The Netherlands.

Mullis, I. V.; Martin, M. O.; Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. TIMSS & PIRLS International Study Center. Boston College, 140 Commonwealth Avenue, Chestnut Hill, MA 02467.

Park, S. P.; Lee, Y. S., & Xing, K. (2016). Investigating the impact of item parameter drift for item Response Theory Models with Mixture Distributions. Distributions. *Frontiers in Psychology,* 7, 255.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*, 207-230.

Risk, N. M. (2015). *The Impact of Item Parameter Drift in Computer Adaptive Testing (CAT).* Doctoral dissertation, University of Illinois at Chicago.

Shultz, K. S., Whitney, D. J., & Zickar, M. J. (2013). *Measurement theory in action: Case studies and exercises*. Routledge.

Swaminathan, H. & Rogers, H. J. (1990). Detection differential item functioning using Logistic regression procedures. *Journal of Educational Measurement*, 27, 361-370

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. *Sage focus editions*, 154, 10-10.

Van der Linden, W. J., & Glas, C. A. (2010). *Elements of adaptive testing*. New York, NY: Springer.

Wei, X. E. (2013). *Impacts of Item Parameter Drift on Person Ability Estimation in Multistage Testing*. Technical Report.

Wells, C. S., Subkoviak, M. J., & Serlin, R. C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement*, *26* (1), 77-87.

Wells, C. S., Hambleton, R. K., Kirkpatrick, R., & Meng, Y. (2014).An examination of two procedures for identifying consequential item parameter drift. *Applied Measurement in Education*, 27, 214–231.

Wu, A. D.; Li, Z.; Ng, S. L., & Zumbo, B. D. (2006). Investigating and comparing the item parameter drift in the mathematics anchor/trend items in TIMSS between Singapore and the United States. In *32nd Annual Conference in International Association for Educational Assessment* (Singapore).

Zumbo, B. D. (1999). *A Handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.