فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی
سال دهم، شماره ۳۰، بهار ۱۳۹۹، صص ۳۱-۷

Journal of Educational Measurement & Evaluation Studies
Vol.10, No. 30, Summer 2020

تغییرناپذیری ساختار عاملی و سؤال‌های آزمون سراسری ریاضی (مورد مطالعه: گروه آزمایشی ریاضی و فنی ۹۶-۱۳۹۵)

مولود علی‌میرزایی*
علی مقدم‌زاده**
اصغر مینائی***
بلال ایزانلو****
کیوان صالحی*****

تاریخ دریافت مقاله: ۹۸/۰۶/۱۳
تاریخ پذیرش مقاله: ۹۸/۱۲/۲۱

چکیده

هدف از اجرای این پژوهش، بررسی تغییرناپذیری ساختار عاملی آزمون سراسری و پارامترهای سؤال (بارعاملی و آستانه) در استان‌ها بود. روش پژوهش توصیفی همبستگی است. برای بررسی هدف پژوهش از هر استان نمونه‌ای به حجم۱۰۰۰ نفر از شرکت‌کنندگان در آزمون ریاضی گروه آزمایشی ریاضی وفنی ۱۳۹۶ انتخاب شد. نتایج نشان داد عملکرد آزمودنی‌ها در تهران، اصفهان، خراسان رضوی، فارس، مازندران، یزد و البرز در بیشتر سؤال‌ها در مقایسه با سایر استان‌ها بهتر است. تعداد سؤال‌های تغییرپذیر در تهران و ایلام از سایر استان‌ها بیشتر است. روش بهینه‌سازی ترازبندی نشان داد ۳۷/۵٪ سؤال‌ها در عامل اول و۱۶٪ درصد در عامل دوم برای تمامی‌استان‌ها دارای تغییرناپذیری تقریبی آستانه‌ها و۸۳٪ سؤال‌ها در عامل اول و ۷۱٪ در عامل دوم دارای تغییرناپذیری بارهای عاملی است. تعداد سؤال‌های تغییرپذیر در عامل دوم، بیشتر از عامل اول است، بنابراین برای مقایسه آزمودنی‌ها در استان‌ها بهتر است از سؤال‌های عامل اول استفاده شود. بررسی DIF در تعداد زیادی از گروه‌ها تنها با روش ترازبندی کافی نیست، اگر گروه خاصی در پژوهش موردنظر باشد لازم است روش‌های DIF دوگروهی استفاده شود.

واژگان کلیدی: تغییرناپذیری اندازه‌گیری، بهینه‌سازی ترازبندی، کارکرد افتراقی چندگروهی، آزمون ورودی دانشگاه.

---

*دانشجوی دکتری سنجش و اندازه‌گیری، دانشگاه تهران، تهران، ایران

**استادیار دانشکده روان‌شناسی و علوم تربیتی دانشگاه تهران، تهران، ایران (نویسنده مسئول: amoghadamzadeh@ut.ac.ir)

*** دانشیار گروه سنجش و اندازه‌گیری، دانشگاه علامه طباطبائی، تهران، ایران

****استادیار گروه برنامه‌ریزی درسی، دانشگاه خوارزمی، تهران، ایران

*****استادیار دانشکده روان‌شناسی و علوم تربیتی، دانشگاه تهران، تهران، ایران

مقدمه

داده‌های حاصل از آزمون‌های سرنوشت‌ساز مانند آزمون سراسری ورودی دانشگاه در ایران، معمولاً برای اطلاع‌رسانی سیاست‌ها و اجرای برنامه‌های درسی و تصمیم‌گیری‌های آموزشی به کاربرده می‌شود. فرض ضمنی استفاده از داده‌های آزمون این است که اندازه‌گیری‌ها در میان استان‌ها و مناطق آموزشی مقایسه‌پذیر هستند. این فرض به معنای مقایسه‌پذیری نمره سؤال‌ها و سازه‌های اندازه‌گیری شده در سنجش‌هاست. هسته اصلی تلاش‌ها و ایجاد تفاسیر دارای روایی بر مبنای نتایج سنجش، نمره‌های مقایسه‌پذیر است. یعنی نمره‌هایی که بتوانند، سازه‌ای یکسان را با مقیاس و درجه عدم قطعیت یکسان در تمامی شرکت‌کنندگان، اندازه‌گیری کنند (اولیوری و ون دیویر[1]، ۲۰۱٤). یکی از دغدغه‌های اصلی آزمون‌سازی این است که آیا استنباط‌های حاصل از نمره‌های آزمون در زیرجامعه‌های مختلف قابل مقایسه است. تحلیل‌های روان‌سنجی مختلفی برای بررسی مقایسه‌پذیری نمره‌ها، مثل بررسی تغییرناپذیری اندازه‌گیری[2] و کارکرد افتراقی سؤال[3] (DIF) اجرا می‌شود (سوتینا و روتکاوسکی[4]، ۲۰۱٤). یک مدل اندازه‌گیری یا تحلیل عاملی تأییدی چندگروهی با نمره‌های مشاهده شده افراد i در گروه j به این صورت است:

$$Y_{ij} = \tau_j + \Lambda_j \eta_{ij} + \varepsilon_{ij} \qquad (۱)$$

که در آن Λ_j و τ_j نشان‌دهنده بارعاملی و عرض از مبدأ گروه j و ε_{ij} و η_{ij} نمره‌های عامل مشترک و باقی‌مانده‌ها برای شخص i در گروه j است. در هر گروه، فرض می‌شود که خطاها دارای توزیع نرمال با میانگین صفر و واریانس Θ_j هستند. اگر تعداد عوامل و سؤال‌ها که روی هر عامل بارگذاری می‌شوند در بین گروه‌ها یکسان باشد، تغییرناپذیری شکلی[5] برقرار است. علاوه بر این، اگر برای همه گروه‌ها $\Lambda_j = \Lambda$

---

1. Oliveri & von Davier
2. Measurement invariance
3. Differential Item Functioning
4. Svetina & Rutkowski
5. configural

تغییرناپذیری متریک[۱] برقرار است. تغییرناپذیری اسکالر[۲] با اضافه کردن برابری $\tau_j = \tau$ در همه گروه‌ها حاصل می‌شود. تغییرناپذیری دقیق[۳] با برابری واریانس باقیمانده‌ها $\Theta_j = \Theta$ حاصل می‌شود که در این مطالعه گنجانده نشده است زیرا تغییرناپذیری اسکالر شرایط کافی برای مقایسه میانگین عوامل بین گروه‌ها را در نظر می‌گیرد (مردیت[۴]، ۱۹۹۳). مقیاسی با تغییرناپذیری اسکار، تغییرناپذیری مناسبی برای بیشتر هدف‌های پژوهشی و عملی دارد. تغییرناپذیری اسکالر نشان می‌دهد که تفاوت در نمره‌های مقیاس به دلیل تفاوت در سطوح واقعی سازه زیربنایی (و نه بنابر سایر دلایل) است. اگر تغییرناپذیری اسکالر برقرار باشد، پژوهشگران می‌توانند میانگین و واریانس عاملی را در میان گروه‌ها مقایسه کنند (بوئن و ماسا[۵]، ۲۰۱۵).

با روش‌های سنتی نمی‌توان آزمون تغییرناپذیری سؤال‌های دو یا چندارزشی را در بیش از دو گروه اجرا کرد. لازم است فرایند آزمون برای هر زوج از گروه‌ها تکرار شود که منجر به تعداد زیادی از آزمون‌های آماری شده که احتمالاً دارای خطاهای استنباط (نوع اول، نوع دوم و بیش برازش) است (فلیک و مک‌کوچ[۶]، ۲۰۱۷). روش ترازبندی[۷] به‌عنوان یک راه جایگزین به منظور ساده‌سازی مراحل خسته‌کننده اندازه‌گیری‌های چندگروهی، برای سازمان‌دهی تغییرناپذیری اندازه‌گیری تقریبی در تعداد زیادی از گروه‌ها معرفی شد (آسپاروهوف و موتن[۸]، ۲۰۱٤). این روش به انتخاب سؤال‌های لنگر، مقایسه برازش مدل‌ها، آزمون‌های متفاوت از مدل‌های آماری به صورت دستی، و مهم‌تر از همه به تعداد مشخصی از گروه‌ها نیاز ندارد و برای موقعیت‌هایی با تعداد بیش از دو گروه ایدئال است (فلیک و مک‌کوچ، ۲۰۱۷). در رویکرد ترازبندی، تغییرناپذیری اندازه‌گیری مفروض نیست و می‌تواند ضمن کشف بهینه‌ترین الگوی تغییرپذیری، میانگین و واریانس عوامل را در هر گروه برآورد کند (آسپاروهوف و موتن، ۲۰۱٤). مشابه چرخش در تحلیل عاملی اکتشافی، بهینه‌سازی ترازبندی برازش مدل تغییرناپذیری

---

1. metric
2. scalar
3. strict
4. Meredith
5. Bowen & Masa
6. Flake & McCoach
7. alignment
8. Asparouhov & Muthén

شکلی را تغییر نمی‌دهد. این روش محدودیت تساوی برآوردهای پارامترها را تحمیل نمی‌کند و اجازه می‌دهد تفاوت در پارامترهای مدل در میان همه گروه‌ها وجود داشته باشد تا بهترین مدل برازش یابد (یعنی مدل تغییرناپذیری شکلی). سپس در فرایند یافتنِ مدل تغییرناپذیری اندازه‌گیری بهینه، میانگین و واریانس عوامل برای هر گروه به وسیلهٔ به حداقل رساندن مقدار تغییرپذیری اندازه‌گیری، محاسبه می‌شود. به دلیل اینکه میانگین و واریانس عوامل هر گروه تابعی از بارهای عاملی و عرض از مبدأ است، هنگامی که مجموع تغییرپذیری اندازه‌گیری در بارهای عاملی و عرض از مبدأها در تمام زوج‌های احتمالی (یعنی تابع زیان کلی که در معادله۲ نشان داده شده است) به حداقل برسد، میانگین‌ها و واریانس‌های عاملی تعیین می‌شوند (کیم و همکاران[1]، ۲۰۱۷).

$$F = \sum_p \sum_{g_m < g_n} w_{g_m g_n} f(\lambda_{pg_m} - \lambda_{pg_n}) + \sum_p \sum_{g_m < g_n} w_{g_m g_n} f(\tau_{pg_m} - \tau_{pg_n}) \quad (۲)$$

که در آن p تعداد نشانگرهای مشاهده شده، g_m و g_n نشان‌دهنده گروه m و n ($n \neq m$) برای هر زوج از گروه‌ها در داده‌ها، λ_{pg_m} و λ_{pg_n} بارهای عاملی گروه m و n و τ_{pg_m} و τ_{pg_n} عرض از مبدأ گروه m و n است. در تابع F برای هر پارامتر اندازه‌گیری، تفاوت بین پارامترهای هر زوج از گروه‌ها از طریق f مقیاس‌سازی شده است که مؤلفه‌ای از تابع زیان است (جنریچ[2]، ۲۰۰۶). عوامل وزنی که نشان‌دهنده اندازه گروه است به صورت زیر تعریف می‌شود:

$$w_{g_m} w_{g_n} = \sqrt{N_{g_m} N_{g_n}} \quad (۳)$$

بر مبنای مؤلفه تابع زیان $f(x) = \sqrt{\sqrt{x^2 + 0.01}}$، هنگامی که تعداد کمی از بارهای عاملی و عرض از مبدأها، تغییرپذیری برجسته داشته باشند و بیشتر بارهای

---

[1]. Kim
[2]. Jennrich

عاملی و عرض از مبدأها دارای تغییرناپذیری تقریبی باشند، تابع F کمینه می‌شود. دو نوع ترازبندی وجود دارد: بهینه‌سازی ثابت وآزاد. در بهینه‌سازی ثابت، میانگین و واریانس عوامل گروه اول با صفر و یک ثابت می‌شوند، در بهینه سازی آزاد، محدودیتی روی میانگین و واریانس عوامل گروه اول وجود ندارد و به آنها به‌عنوان پارامترهای اضافی که باید برآورد شود، نگاه می‌شود (کیم و همکاران، ۲۰۱۷).

مرور پیشینه نشان داد که تاکنون در ایران در رابطه با تغییرناپذیری آزمون ورودی دانشگاه‌ها در استان‌ها، پژوهشی انجام نگرفته است. تنها چند مورد پژوهش در ارتباط با کارکرد افتراقی سؤال در دو گروه و تنها برحسب یک ویژگی و آن هم بیشتر به لحاظ جنسیت انجام گرفته است. به‌عنوان مثال، کارکرد افتراقی جنسیتی با استفاده از روش مانتل- هنزل و روش مبتنی بر نظریه پرسش- پاسخ برای آزمون ریاضی گروه آزمایشی ریاضی سال ۸۲ بررسی شده است (شریفی یگانه، ۱۳۹۱). در پژوهشی دیگر، سهمیه‌بندی منطقه یک و دو به‌عنوان شاخصی برای پایگاه اجتماعی و اقتصادی، جنسیت و استان اخذ دیپلم (دو استان به منظور بررسی تفاوت زبان تکلم انتخاب شده است، اصفهان و آذربایجان) در برخی درس‌های تخصصی رشته‌های مختلف آزمون سراسری در سال‌های ۱۳۸۷ تا ۱۳۹۰ در بررسی کارکرد افتراقی سؤال مد نظر قرار گرفته است و از رویکرد تحلیل رگرسیون لوجستیک دووجهی استفاده شده است (گرامی‌پور و همکاران، ۱۳۹۵). همچنین کارکرد افتراقی سؤال و آزمون، مرتبط با جنسیت در آزمون‌های تخصصی یک دفترچه آزمون در پنج گروه آزمایشی شرکت‌کننده در آزمون‌های سراسری سال های ۱۳۸۷ و ۱۳۹۰بررسی شده و برای مطالعه DIF از تحلیل رگرسیون لوجستیک و برای مطالعه DTF از رویکرد مبتنی بر IRT استفاده شده است (گرامی‌پور و همکاران، ۱۳۹٦). ولی تاکنون پژوهشی به منظور بررسی مقایسه‌پذیری نمره‌ها در بیشتر از دو گروه انجام نگرفته است. در مقایسه‌پذیری نمره‌های آزمون ورودی دانشگاه‌ها در استان‌های ایران که تعداد گروه‌ها زیاد است، ضروری است روش‌های مناسب آماری اتخاذ شود.

هر سال پس از اعلام نتایج کنکور، عملکرد آزمودنی‌ها در استان‌های کشور مقایسه می‌شود؛ به‌عنوان مثال سرپرست وزارت آموزش‌وپرورش در ۲۰ مرداد ماه ۱۳۹۸ با طرح این پرسش که آیا عدالت آموزشی را به خوبی پیاده می‌کنیم، اظهار داشت که بیش از ۸۰ درصد کل پذیرفته‌شدگان کنکور به ۵ تا ٦ استان اختصاص دارد (روزنامه دنیای اقتصاد، ۳۵۵۸۹۳۱). سلیمی و پاسالاری (۱۳۹٦) در پژوهش خود با عنوان «بررسی نقش

ویژگی‌های اجتماعی و تحصیلی دانش‌آموختگان دبیرستانی استان هرمزگان در موفقیت آنان در آزمون ورودی دانشگاه‌ها»، این مبحث را مطرح کرد که تعداد پذیرفته‌شدگان در آزمون ورودی دانشگاه‌ها، نشانگر میزان دست‌یابی و بهره‌مندی افراد از آموزش عالی است. در این پژوهش بیان شده است که عدالت آموزشی در ایران، کاستی‌هایی دارد و عامل مهم در برخورداری برخی گروه‌های جمعیتی از آموزش عالی با کیفیت‌تر است. مقایسه درصد رتبه‌های برتر یا درصد پذیرفته‌شدگان در هر استان، اطلاعات توصیفی ارائه می‌کند اما نشان نمی‌دهد که تفاوت‌های عملکردی آزمودنی‌ها اصولًا مربوط به مهارت‌ها و توانایی‌های تحت سنجش، یا عوامل نامرتبط است. تفاوت، هنگامی معنی‌دار است که توانایی مورد نظر در گروه‌ها کنترل شده باشد، یعنی تفاوت در عملکرد آزمودنی به‌تنهایی نمی‌تواند نشان‌دهنده سوگیری سؤال یا آزمون باشد؛ زیرا ممکن است آزمودنی‌ها در استان‌ها واقعاً در توانایی مورد نظر تفاوت داشته باشند بنابراین لازم است کارکرد افتراقی سؤال‌ها به‌عنوان گواهی تجربی، برای وجود یا نبود سوگیری سؤال، بررسی شود که این امر گامی مهم در روایی سنجش است.

بنابراین به منظور مقایسه یا رتبه‌بندی استان‌ها بر مبنای عملکرد آزمودنی‌ها در آزمون سراسری، لازم است بررسی شود که آیا آزمون، سازه‌ای یکسان را در میان استان‌ها اندازه می‌گیرد. سپس لازم است سؤال‌ها به منظور وجود کارکرد افتراقی در میان استان‌ها بررسی شوند. اما اگر یافته‌های توصیفی یا بحث و اختلاف نظر در مورد عملکرد آزمودنی‌ها در استان خاصی مطرح باشد، مثلاً، استان یزد بالاترین میزان و استان هرمزگان کمترین میزان پذیرفته‌شدگان را دارد (سلیمی و پاسالاری، ۱۳۹٦)، لازم است استان موردنظر با سایر استان‌ها به صورت دوبه‌دو مقایسه شود و دقیقاً مشخص شود کدام سؤال‌ها باعث کارکرد افتراقی شده‌اند. بنابراین هدف‌های مورد بررسی در این پژوهش عبارت‌اند از: ۱- بررسی ساختار عاملی آزمون ریاضی گروه آزمایشی ریاضی ۱۳۹٦؛ ۲-مقایسه‌پذیری ساختار عاملی سؤال‌های آزمون ریاضی گروه آزمایشی ریاضی ۱۳۹٦ در استان‌های ایران بر مبنای رویکرد تحلیل عاملی؛ ۳- مقایسه میانگین عوامل در میان استان‌ها و رتبه‌بندی و گروه‌بندی استان‌های ایران بر اساس میانگین عوامل زیربنایی؛ ٤- شناسایی سؤال‌های دارای تغییرپذیری اندازه‌گیری با روش چندگروهی بهینه‌سازی ترازبندی.

روش

پژوهش‌های توصیفی شامل مجموعه روش‌هایی است که هدف آنها توصیف شرایط یا پدیده‌های مورد بررسی است و اجرای پژوهش توصیفی می‌تواند صرفاً برای شناخت بیشتر شرایط موجود و یاری دادن به فرایند تصمیم‌گیری باشد (سرمد و همکاران، ۱۳۸٤). از این‌رو، این پژوهش به‌طور عام جزو پژوهش‌های توصیفی (غیرآزمایشی) است. روش پژوهش حاضر از حیث نوع تجزیه‌وتحلیل‌های آماری جزو دسته تحلیل‌های عاملی در تحقیقات همبستگی است. جامعه پژوهش شامل همه شرکت‌کنندگان گروه ریاضی در آزمون سراسری سال ۱۳۹٦ است. در پژوهش حاضر ماده امتحانی ریاضی شامل ۵۵ سؤال چهارگزینه‌ای از آزمون اختصاصی گروه آزمایشی ریاضی و فنی برای بررسی تغییرناپذیری اندازه‌گیری در میان استان‌های ایران مورد استفاده قرار گرفت. با توجه به اینکه سازمان سنجش آموزش کشور نتایج آزمون تمامی شرکت‌کنندگان را در اختیار پژوهشگران قرار نمی‌دهد، بنابراین کل جامعه در دسترس نیست. شواهد موجود در پیشینه همگی بر نمونه‌هایی دست کم به حجم ۱۰۰۰ نفر توافق دارند (ایزانلو و همکاران، ۱۳۹۳). بنابراین به‌طور تقریبی نمونه‌ای به حجم ۱۰۰۰ نفر از هر استان انتخاب شد.

آزمون سراسری ورودی دانشگاه‌ها چهارگزینه‌ای است که به پاسخ‌های غلط نمره منفی اختصاص می‌یابد. بنابراین آزمودنی‌ها تشویق می‌شوند به سؤال‌هایی که از پاسخ آن اطمینان ندارند یا پاسخ را نمی‌دانند، پاسخ ندهند. نتایج پژوهش چگینی و همکاران (۱۳۹۸) نشان داد هنگامی‌که با پاسخ‌های گمشده به‌عنوان غلط برخورد شود، برای افراد با الگوی پاسخ مختلف و تعداد پاسخ صحیح یکسان، برآورد توانایی یکسانی به دست می‌آید. همچنین درنظرگرفتن گمشده‌ها به‌عنوان غلط، باعث افزایش دشواری سؤال‌ها شده و آسان‌ترین و دشوارترین سؤال‌ها با هنگامی‌که گمشده‌ها دستکاری نشود، متفاوت است. در این پژوهش با پاسخ‌های گمشده به‌عنوان پاسخ غلط برخورد شد. از آنجایی‌که هدف مقایسه عملکرد آزمودنی‌ها در استان‌هاست و در تمامی استان‌ها رویکرد یکسانی برای برخورد با گمشده‌ها اتخاذ شده است، بیش‌برآورد دشواری نمی‌تواند به سود یا زیان استانی باشد. البته روشی برای برخورد با داده‌های گمشده وجود ندارد که در همه موارد برتر باشد (میسلوی و وو[1]، ۱۹۹٦)، بر این اساس، لازم است در پژوهش‌های

[1]. Mislevy & Wu

آینده، برای انتخاب بهترین راهبرد برای رسیدگی به داده‌های گمشده در هنگام مقایسه گروه‌ها مطالعه و بررسی شود.

به دلیل مشخص نبودن ساختار عاملی آزمون، ابتدا تعداد عامل‌ها با استفاده از روش آماری ناپارامتری مبتنی بر فرض استقلال اساسی با استفاده از نرم‌افزار DIMPACK (استوت و همکاران[1]، ۲۰۰۱) و آزمون MAP با استفاده از بسته psych (رول[2]، ۲۰۱۵) در نرم افزار R انجام گرفت. ساختار عاملی با رویکرد تحلیل عاملی غیرخطی در نرم‌افزار NOHARM4 (فراسر و مک‌دونالد[3]، ۱۹۸۸) بررسی شد. پس از مشخص کردن ساختار عاملی آزمون، مدل پایه مناسب برای برازش به داده‌ها شناسایی شد و بر اساس این مدل، تغییرناپذیری اندازه‌گیری آزمون ریاضی در استان‌ها با استفاده از تابع cfa در بسته lavaan بررسی شد (روسیل[4]، ۲۰۱۲). سپس بر اساس روش ترازبندی با نرم‌افزار Mplus نسخه ۷/٤ (موتن و موتن، ۲۰۱۷–۱۹۹۸)، میانگین عامل‌ها در استان‌ها مقایسه شد.

نتایج

از آنجایی که ساختار عاملی آزمون‌های ورودی دانشگاه از قبل مشخص نیست، ساختار عاملی آزمون ریاضی گروه آزمایشی ریاضی‌وفنی ۱۳۹٦ برای هدف اول پژوهش، بررسی شد. نخست، تک‌بعدی بودن داده‌ها با استفاده از روش DIMTEST مورد آزمون قرار گرفت. مقدار آماره آزمون T برابر ۱/۱٤۹ با سطح معنی‌داری ۰/۱۲۵ به دست آمد که حاکی از تک‌بعدی بودن آزمون است. بر اساس پیشنهاد متخصصان، گام بعدی استفاده از روش DETECT برای تعیین تعداد ابعاد است (ایزانلو و همکاران، ۱۳۹۳). مقادیر نزدیک به صفر شاخص DETECT حاکی از تک‌بعدی بودن و مقادیر نزدیک یک یا بزرگ‌تر از آن به چندبعدی بودن اشاره دارد. مقدار DETECT حدود ۰/۳۷ بود که تک‌بعدی بودن را نشان نمی‌دهد. در حالی که آزمون DIMTEST تک‌بعدی بودن داده‌ها را نشان داد این گونه تناقض‌ها در نتایج مربوط به داده‌های واقعی بعید نیست. در شرایطی که ساختار ساده به دلایلی مثل همبستگی عامل‌ها یا پیچیدگی ساختار نقض

---

1. Stout et al
2. Revelle
3. Fraser & McDonald
4. Rosseel

شود نتایج حاصل DETECT و DIMTEST به‌راحتی تفسیر نمی‌شود (ایزانلو و همکاران، ۱۳۹۳).

نتایج آزمون کمینه میانگین همبستگی‌های تفکیکی[1] (MAP) براساس همبستگی تتراکوریک مناسب با داده‌های دوارزشی، نشان داد، MAP کمترین مقدار خود را در ۲ عامل اختیار می‌کند. این روش در هر یک از گام‌های متوالی متوسط همبستگی‌های غیرقطری مجذور شده موجود در یک ماتریس را ملاک انتخاب تعداد ابعاد قرار می‌دهد. استخراج عامل‌ها تا جایی ادامه می‌یابد که متوسط همبستگی تفکیکی مجذور شده غیرقطری به کمترین مقدار برسد (ایزانلو و همکاران، ۱۳۹۳).

به منظور یافتن مدل مناسب برای برازش به داده‌ها و ترتیب قرار گرفتن سؤال‌ها در عوامل، از روش وارسی روایی[2] استفاده شد، یعنی ابتدا داده‌ها به دو مجموعه داده تقسیم شد، در مجموعه داده اولیه تحلیل عاملی اکتشافی به ترتیب با دو عامل (به دلیل آنکه نتایج روش‌های بررسی ابعاد زیربنایی، نشان از وجود ۲ عامل دارد) انجام گرفت و روایی مدل به دست آمده در این مجموعه داده، با استفاده از تحلیل عاملی تأییدی در مجموعه داده دوم سنجیده شد (دیانا و توماسی[3]، ۲۰۰۲). نتایج حاصل از چرخش پروماکس برای دو عامل نشان داد که به‌جز برخی سؤال‌ها که دارای ساختار ساده نیستند، سؤال‌های ۱ تا ٢٤ در عامل اول و سؤال‌های ۲۵ تا ۵۵ در عامل دوم، دارای بارعاملی بالایی هستند. جدول (۱) بارهای عاملی حاصل از چرخش پروماکس را نشان می‌دهد. به منظور بررسی روایی، مدل‌های یک‌عاملی و دوعاملی به داده‌ها برازش داده شد، شاخص‌های برازش CFI، TLI، RMSEA، SRMR برای مدل یک‌عاملی ۰/۹۲۰، ۰/۹۱۷، ۰/۰۳۹، ۰/۰٦٤ و برای مدل دوعاملی ۰/۹۵۵، ۰/۹۵۳، ۰/۰۲۹، ۰/۰۵، نشان می‌دهد مدل دوعاملی دارای برازش بهتری است.

تحلیل محتوای سؤال‌ها توسط ٤ مدرس ریاضی نشان داد که سؤال‌های ۱ تا ٢٤ مربوط به مبحث حساب، دیفرانسیل و انتگرال است. این سؤال‌ها نیازمند محاسبات جبری بوده و از سؤال‌هایی هستند که نسبت پاسخ‌های درست آنها تقریباً زیاد است. سؤال‌های ۲۵ تا ۳۲ مربوط به هندسه پایه است. سؤال‌های ۳۳ تا ۳٦ هندسه تحلیلی، سؤال‌های ٤۱ و ٤۲ آمار و مدل‌سازی، ٤۷ و ٤۸ مربوط به احتمال است. سؤال‌های ۳۷

---

[1]. minimum average partial correlation
[2]. Cross-validation
[3]. Diana & Tommasi

تا ٤٠ مربوط به جبرخطی، سؤال‌های ٤٣ تا ٤٦ جبرواحتمال و سؤال‌های ٤٩ تا ٥٥ مربوط به ریاضیات گسسته است. بنابراین عامل اول مربوط به سؤال‌های حساب دیفرانسیل و انتگرال و عامل دوم سؤال‌های هندسه و جبرواحتمال و ریاضیات گسسته است.

جدول (۱) بارهای عاملی حاصل از چرخش پروماکس

سؤال	عامل ۱	عامل ۲	سؤال	عامل ۱	عامل ۲
۱	۰/٤۱٥	۰/۳۱۰	۲۹	-۰/۱۳٦	۰/۷٦٦
۲	۰/٥۱۹	۰/۰۸۳	۳۰	۰/۱۸۲	۰/٦٦۳
۳	۰/٦۳۳	۰/۱۳۸	۳۱	-۰/۱٦۹	۰/٥۹۳
٤	۰/٦۸۰	۰/۱۰٤	۳۲	-۰/۰٤۲	۰/۷٥۲
٥	۰/٥٤٦	۰/۳۲۹	۳۳	۰/٤۱۲	۰/٤۱۸
٦	۰/٦۸۲	۰/۰۷۷	۳٤	۰/۲۲٥	۰/٦۰۳
۷	۰/٤٦۰	۰/۲٤٥	۳٥	۰/۲۲۹	۰/٥۰۸
۸	۰/۷٦٦	۰/۰۱٥	۳٦	۰/۲٥٤	۰/٤۷۷
۹	۰/٥۷٥	۰/۱۰۹	۳۷	۰/٤۱۷	۰/۳٦٥
۱۰	۰/۳٦٤	۰/۳۸۷	۳۸	۰/۳۱۲	۰/٤٤۲
۱۱	۰/۸۲٤	-۰/۰۷۱	۳۹	۰/٤۱۹	۰/۳۷۸
۱۲	۰/۹۱٦	-۰/۰۷٥	٤۰	۰/۳٦٤	۰/٤۲۱
۱۳	۰/٥۰٤	-۰/۰۳۱	٤۱	۰/۰۹۸	۰/٤۲٦
۱٤	۰/۷٥٥	۰/۰٥۸	٤۲	۰/۱٤۰	۰/٦۳٥
۱٥	۰/٦۳۷	۰/۲۰۰	٤۳	۰/۱۳۸	۰/٦۰٦
۱٦	۰/۷۰٥	۰/۱۲۱	٤٤	۰/۲۰۷	۰/٥۷۹
۱۷	۰/٦۲٦	۰/۲۰٤	٤٥	۰/۲٦۳	۰/٤۳۷
۱۸	۰/۹۱۲	-۰/۰٤۰	٤٦	۰/۰۹٥	۰/٥٥۹
۱۹	۰/۷۷۳	۰/۱۲۸	٤۷	۰/۰۰۸	۰/٥۳٦
۲۰	۰/۸۷٥	-۰/۰۷۲	٤۸	۰/۰٦۷	۰/٥۹۰
۲۱	۰/٥۱۰	۰/۳٤۳	٤۹	-۰/۰۱۰	۰/٦۲٤
۲۲	۰/۳۷۲	۰/۳٤۷	٥۰	۰/۰٤۰	۰/٦٥٤
۲۳	۰/٦۲۲	۰/۲٦٥	٥۱	۰/۱۸۳	۰/٦۲۳
۲٤	۰/٦٤۷	۰/۰۸۳	٥۲	-۰/۰۳٦	۰/٦۰٦
۲٥	۰/۰۱٤	۰/٦۸۷	٥۳	۰/۰۳۳	۰/٦۷۷
۲٦	-۰/۰۹۹	۰/۷٤۸	٥٤	۰/۲۸۲	۰/٦۰۸
۲۷	۰/۱۲٥	۰/٦۸۲	٥٥	۰/۳۱٤	۰/٤۰۲
۲۸	۰/۳٥۱	۰/۷۰۰			

برای ارزیابی مدل‌های تحلیل عاملی تأییدی، شاخص‌های برازش مدل با نقاط برش مثل شاخص CFI>۰/۹۵، RMSEA<۰/۰۶، SRMR<۰/۰۸، معمولاً مورد استفاده قرار می‌گیرد (هو و بنتلر[1]، ۱۹۹۹). از آنجا که در تحلیل چندگروهی هنگامی که تعداد گروه‌ها زیاد است (بیشتر از ۱۰ گروه) ، مقدار RMSEA صرف نظر از مدل واقعی بیشتر از ۰/۰۵ می‌شود، روتکاوسکی و سوتینا[2] (۲۰۱٤) نقطه برش آزادانه‌تر ۰/۱ برای RMSEA در ۱۰ یا ۲۰ گروه پیشنهاد کردند. شاخص‌های برازش مدل دوعاملی به داده‌ها عبارت است از CFI برابر ۰/۹۵۵ و RMSEA برابر ۰/۰۲۹ که نشان‌دهنده برازش مناسب مدل است. بنابراین مدل دوعاملی به‌عنوان مدل پایه در نظر گرفته شد و تغییرناپذیری اندازه‌گیری این مدل، در میان ۳۱ استان ایران بررسی شد.

به منظور بررسی هدف دوم، مقایسه‌پذیری ساختار عاملی سؤال‌های ریاضی آزمون سراسری سال ۱۳۹٦ در استان‌های ایران بر مبنای رویکرد تحلیل عاملی، تغییرناپذیری شکلی، متریک و اسکالر مدل پایه، بررسی شد. برای مقایسه مدل‌های تحلیل عاملی تأییدی چندگروهی، به دلیل اینکه آماره کای دو هنگامی که مدل دقیقا مناسب نیست، به اندازه نمونه حساس است (بنتلر و بونت[3]، ۱۹۸۰)، تعدادی شاخص مناسب جایگزین برای ارزیابی برازش در آزمون تغییرناپذیری اندازه‌گیری پیشنهاد شده است. چئونگ و رنسولد[4] (۲۰۰۲) تغییر در مقدار CFI کمتر از ۰/۰۱ را به‌عنوان شاهدی بر تغییرناپذیری معرفی کردند. چن[5] (۲۰۰۷) افزود که RMSEA باید کمتر از ۰/۰۱۵ باشد. با این حال، این پیشنهادها برای تعداد گروه‌های کم (مثلا دو گروه) است. وقتی که تعداد گروه‌ها زیاد باشد (۱۰ یا ۲۰) روتکاوسکی و سوتینا (۲۰۱٤) معیارهای آزادانه‌تری مثل $\Delta CFI \leq ۰/۰۲$ و $\Delta RMSEA \leq ۰/۰۳$ را برای آزمون تغییرناپذیری متریک و $\Delta CFI \leq ۰/۰۱$ و $\Delta RMSEA \leq ۰/۰۱۵$ را برای آزمون تغییرناپذیری اسکار پیشنهاد کردند.

---

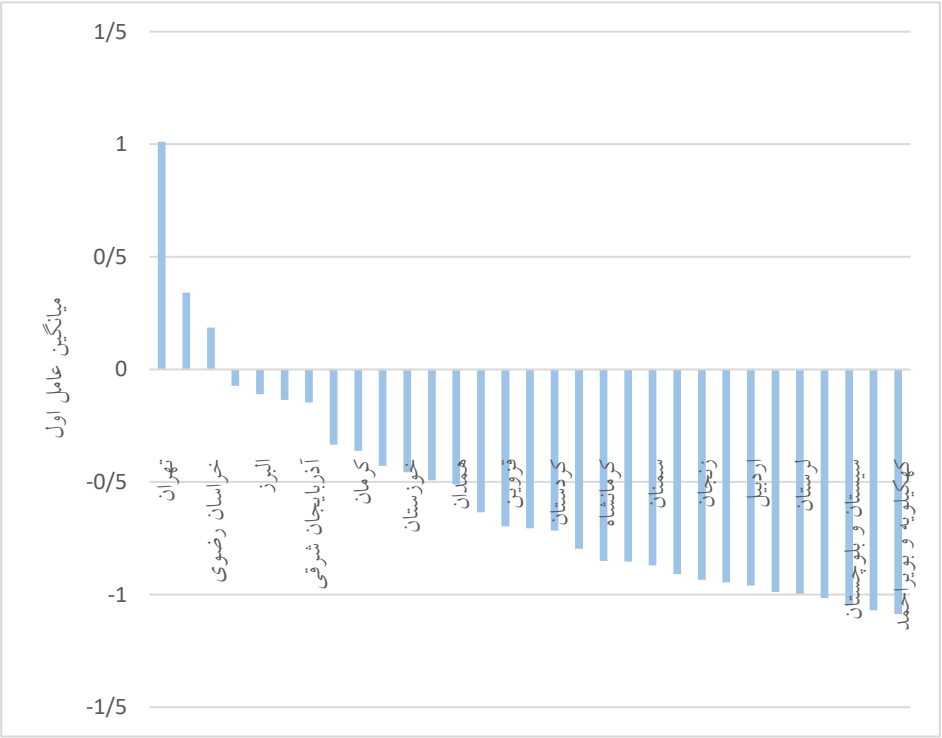
1. Hu & Bentler
2. Rutkowski & Svetina
3. Bentler & Bonnet
4. Cheung & Rensvold
5. Chen

جدول (۲) شاخص‌های برازش مدل دوعاملی در تغییرناپذیری شکلی، متریک و اسکالر

	CFI	ΔCFI	RMSEA	$\Delta RMSEA$	SRMR
شکلی	۰/۹۵٤	–	۰/۰۲۱	–	۰/۰۹۳
متریک	۰/۹٦۳	-۰/۰۰۹	۰/۰۱۸	-۰/۰۰۳	۰/۱۱۰
اسکالر	۰/۹۵۹	۰/۰۰٤	۰/۰۱۹	۰/۰۰۱	۰/۰۹٦

جدول (۲) شاخص‌های آزمون تغییرناپذیری اندازه‌گیری مدل دو عاملی را به داده‌ها نشان می‌دهد. ΔCFI و ΔRMSEA برای این مدل، از نقاط برش کمتر است بنابراین فرض تغییرناپذیری اندازه‌گیری اسکالر رد نمی‌شود. اما هنگام استفاده از نمونه‌های بزرگ (بزرگ‌تر از ۵۰۰) تحلیل عاملی تأییدی چندگروهی همواره مدل اسکالر را نشان می‌دهد. همچنین در تحلیل عاملی تأییدی، آماره‌های مطلق برازش مثل کای دو، مدل را در نمونه‌های بزرگ رد می‌کند (بالن[1]، ۱۹۹۰). برای رفع این مشکلات و بررسی دقیق‌تر اینکه کدام سؤال‌ها منابع تغییرناپذیری هستند و کدام گروه‌ها بیشترین مشکلات اندازه‌گیری را به وجود می‌آورند. همچنین به منظور مقایسه میانگین عوامل در میان استان‌ها برای بررسی سایر هدف‌های پژوهش از روش بهینه‌سازی ترازبندی استفاده شد.

[1]. Bollen

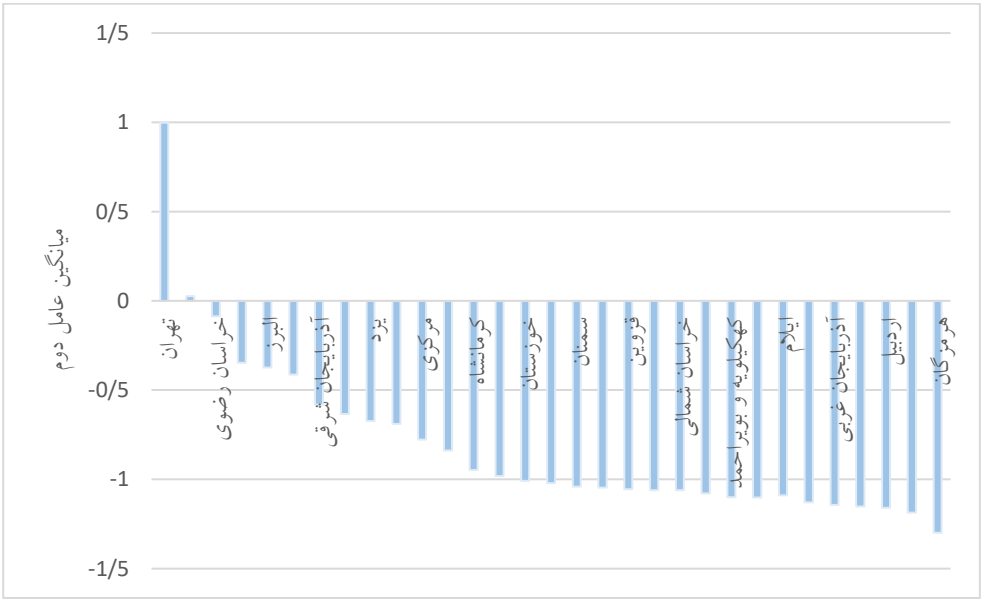


شکل (۱) میانگین عامل اول به تفکیک استان‌ها در آزمون ریاضی

برای بررسی هدف سوم پژوهش، مقایسه میانگین عوامل در میان استان‌ها و رتبه‌بندی و گروه‌بندی استان‌های ایران بر اساس میانگین عوامل زیربنایی، استان‌ها به ترتیب در اندازه میانگین اولین و دومین عامل در آزمون ریاضی از بزرگ به کوچک با روش ترازبندی مرتب شدند. شکل‌های (۱ و ۲) میانگین‌های عاملی در استان‌ها در عامل اول و دوم است که نشان می‌دهد میانگین عاملی در تهران با سایر استان‌ها تفاوت زیادی دارد و به‌طور کلی میانگین عاملی در عامل اول از عامل دوم بیشتر است.

آزمون معنی‌داری تفاوت میانگین‌های عاملی در گروه‌ها در روش ترازبندی به این صورت است؛ در گام اول مجموعه گروه‌های تغییرناپذیر ابتدایی برای هر پارامتر (آستانه و بارعاملی) تعیین می‌شود. همه گروه‌ها به صورت دوبه‌دو با آزمون t مقایسه می‌شوند، گروه‌هایی که دارای سطح معنی‌داری بزرگتر از ۰/۰۱ هستند در یک مجموعه قرار می‌گیرند. سپس بزرگ‌ترین مجموعه از گروه‌ها به‌عنوان مجموعه ابتدایی انتخاب

می‌شود. ابتدا میانگین پارامتر در این مجموعه محاسبه شده، سپس برای هر گروه، یک آزمون برای مقایسه مقدار پارامتر با میانگین مجموعه ابتدایی، اجرا می‌شود. هر گروهی که دارای سطح معنی‌داری بالاتر از ۰/۰۰۱ بود، به گروه ابتدایی اضافه می‌شود. این روش تا جایی تکرار می‌شود که مجموعه تغییرناپذیر، پایدار شود و هیچ گروهی به مجموعه تغییرناپذیر اضافه نشود. نرخ خطای نوع اول در روش ترازبندی به وسیله تنظیم مقدار ملاک به ۰/۰۰۱ کنترل شده است (آسپاروهوف و موتن، ۲۰۱٤).



شکل (۲) میانگین عامل دوم به تفکیک استان‌ها در آزمون ریاضی

با توجه به تفاوت معنی‌دار بین استان‌ها در میانگین عاملی، استان‌ها گروه‌بندی شدند، که این اطلاعات در جدول (۳) قرار دارد. هر گروه در جدول (۳) نشان‌دهنده استان‌هایی است که در عامل اول و دوم تفاوت معنی‌داری با یکدیگر ندارند. بنابراین در عامل اول، میانگین عاملی برای ٤ استان مازندران، البرز، فارس، آذربایجان شرقی تفاوت معنی‌داری نداشت. جدول (۳) توجه برنامه‌ریزان و مسئولان آموزشی را به وضعیت منطقه‌ای موجود جلب می‌کند تا برنامه‌ریزی‌های کوتاه‌مدت و بلندمدت، تخصیص بهینه منابع و در نهایت کاهش اختلاف بین مناطق آموزشی صورت گیرد.

جدول (۳) گروه‌بندی استان‌ها بر اساس نداشتن تفاوت معنی‌دار

گروه‌ها	استان‌های بدون تفاوت معنی‌دار در عامل اول	استان‌های بدون تفاوت معنی‌دار در عامل دوم
۱	تهران	تهران
۲	اصفهان	اصفهان، خراسان رضوی
۳	خراسان رضوی	فارس، البرز، مازندارن
٤	مازندران، البرز، فارس، آذربایجان شرقی	آذربایجان شرقی، قم، یزد، کرمان، مرکزی
٥	یزد، کرمان، قم	همدان، کرمانشاه، خراسان جنوبی
٦	خوزستان، مرکزی، همدان	خوزستان، کردستان، سمنان، گیلان، قزوین، زنجان، خراسان شمالی، لرستان، کهکیلویه‌و بویر احمد، گلستان، ایلام، بوشهر، آذربایجان غربی، چهارمحال‌وبختیاری، اردبیل، سیستان و بلوچستان
۷	آذربایجان غربی، قزوین، گیلان، کردستان	هرمزگان
۸	چهارمحال‌وبختیاری، کرمانشاه، خراسان جنوبی، سمنان، ایلام	
۹	زنجان، گلستان، اردبیل، هرمزگان، لرستان، خراسان شمالی، سیستان و بلوچستان، بوشهر، کهکیلویه‌وبویراحمد	

بررسی هدف چهارم، شناسایی سؤال‌های دارای تغییرپذیری اندازه‌گیری با استفاده از روش چندگروهی بهینه‌سازی ترازبندی نشان داد تعداد ١٤(٢٦٪) سؤال، تغییرناپذیری اندازه‌گیری آستانه در تمامی استان‌ها دارند، شامل سؤال‌های ۱، ۲، ٥، ٦، ۱۱، ١٦، ۱۷، ۱۹، ٢٤، ۳۲، ۳۳، ٤٢، ٥۳، ٥٤ که برای تمامی استان‌ها دارای تغییرناپذیری اندازه‌گیری تقریبی آستانه‌ها (۳۷/٥ درصد عامل اول، ١٦ درصد عامل دوم) هستند. تعداد ۲۰ سؤال از ٢٤ سؤال عامل اول (۸۳٪) و تعداد ۲۲ سؤال از ۳۱ سؤال عامل دوم (۷۱٪) دارای تغییرناپذیری اندازه‌گیری بارهای عاملی هستند. بنابراین ٤۲ سؤال از ٥٥ سؤال (٧٦٪) تغییرناپذیری اندازه‌گیری متریک دارند. به‌طور کلی، تعداد ۱۹ استان از ۳۱ استان در سؤال‌های آزمون ریاضی دارای تغییرپذیری هستند، تعداد سؤال‌های دارای تغییرپذیری بارعاملی و آستانه در هر استان در جدول (٤) قرار دارد. دو استان ایلام و تهران در تعداد بیشتری از سؤال‌ها دارای تغییرپذیری اندازه‌گیری هستند.

جدول (٤) تعداد سؤال‌های دارای تغییرپذیری در هر استان

استان	بارعاملی	آستانه	استان	بارعاملی	آستانه	استان	بارعاملی	آستانه
ایلام	۰	۱۱	سیستان	۰	۵	قم	۰	۳
تهران	۷	٦	خراسان رضوی	۲	۵	خراسان جنوبی	۰	۲
فارس	۰	٤	مازندران	۰	٤	مرکزی	۰	۲
اصفهان	۳	۲	خوزستان	۰	٤	قزوین	۰	۱
کهکیلویه وبویراحمد	۰	۵	هرمزگان	۳	۱	کردستان	۰	۱
البرز	۰	۵	یزد	۰	٦	کرمان	۰	۱
						آذربایجان‌شرقی	۰	۳

شاخص R^2 در روش ترازبندی برای پارامترهای آستانه و بار عاملی هر سؤال محاسبه می‌شود که نشان‌دهنده میزان تغییرپذیری پارامتر اندازه‌گیری در مدل شکلی (آستانه و بار عاملی) در میان استان‌ها است که توسط میانگین و واریانس عوامل تبیین می‌شود. محدوده این شاخص بین صفر و یک است و هر اندازه به عدد یک نزدیک‌تر باشد، پارامتر تغییرناپذیرتر است. سؤال‌های ۵، ١٦، ۱۱، ٤ به ترتیب سؤال‌هایی با بالاترین شاخص R^2 (جدول۵)، یعنی تغییرناپذیرترین سؤال‌ها در پارامتر آستانه و بنابراین سؤال‌های مناسبی برای مقایسه آزمودنی‌ها در استان‌ها هستند. بیشتر سؤال‌ها در پارامتر بار عاملی دارای شاخص R^2 بسیار پایین یا صفر هستند. به‌عنوان مثال در سؤال‌های ۱۷، ۲۰، ۲۱ شاخص R^2 برابر صفر است در صورتی که هیچ استانی دارای تغییرپذیری برای این سؤال‌ها مشخص نشده است. این موارد هنگامی اتفاق می‌افتد که توان کافی برای ایجاد تغییرپذیری وجود ندارد، مثلاً اندازه نمونه کوچک یا تعداد داده‌های گمشده برای سؤال زیاد است یا به دلیل خانه‌های خالی در جدول‌های دوطرفه سؤال‌ها، مقدار انحراف استاندارد خیلی بزرگ است یا هنگامی که بار عاملی سؤال به صفر نزدیک است. در نمونه مورد بررسی برای این پژوهش سؤال‌های ۲۲، ٢٦، ٣٤، ۳۵، ٣٦، ٤٥، ٤٦، ٤٧، ٥٢، ۵۳ دارای خانه‌های خالی در جدول‌های دوطرفه هستند. جدول دو طرفه سؤال ۲۲ و ۳۲ دارای خانه خالی است یعنی فردی در نمونه وجود ندارد که هم به سؤال ۲۲ و هم به سؤال ۳۲ پاسخ درست داده باشد. این ده سؤال با بیشتر سؤال‌ها دارای خانه خالی در جدول‌های دوطرفه هستند. برای پرهیز از این

مشکلات لازم است در بررسی‌های بعدی تعداد نمونه بیشتری انتخاب شود. به‌طور کلی تعداد ۲۱ استان از ۳۱ استان در سؤال‌ها دارای تغییرپذیری هستند، دو استان ایلام و تهران در تعداد بیشتری از سؤال‌ها دارای تغییرپذیری اندازه‌گیری هستند. مقادیر شاخص R^2 و سهم پارامترِ هر سؤال در تابع زیان در جدول (۵) نشان داده شده است. سهم هر پارامتر در تابع زیان، مقادیری است که برای هر پارامتر به تابع زیان افزوده می‌شود. پارامتر هر سؤالی که کمترین مشارکت را در تابع زیان داشته باشد، سؤال مناسب‌تری از لحاظ تغییرناپذیری در میان استان‌ها است. سؤال‌های ۵، ۱۲ و ۱۸ کمترین سهم را در تابع زیان برای پارامتر آستانه دارند، سؤال ۱۱، ۱٤، ۱٦، ۱۸ کمترین سهم را در تابع زیان برای پارامتر بارعاملی دارند. به‌طور کلی، با توجه به جدول (۵) از نظر هر دو شاخص سؤال‌های عامل اول مناسب‌تر هستند.

جدول (۵) شاخص R^2 و سهم تابع زیان برای پارامترهای بارعاملی و آستانه‌های سؤال‌های آزمون ریاضی

سؤال	R^2بارعاملی	سهم بارعاملی درتابع زیان	R^2آستانه	سهم آستانه در تابع زیان	سؤال	R^2بارعاملی	سهم بارعاملی در تابع زیان	R^2آستانه	سهم آستانه در تابع زیان
۱	۰/۲۱٦	−۲۰٦/۸۵۹	۰/۸۰۹	−۲۰٤/۷۳٦	۲۹	۰/۱۰۳	−۲۸۹/۳۳٦	۰/۰۰۰	−۳۲۰/۰٦۷
۲	۰/۳۳۱	−۱۹۷/۱۷٤	۰/۹٤۰	−۲۰۲/۱۵۸	۳۰	۰/۲۷۱	−۱۹۸/۹٦٦	۰/۹۰٦	−۱۸۹/٦۱۰
۳	۰/۲۵٤	−۲۰۳/۲۷۵	۰/۹٤۷	−۱۷۷/۸۷۰	۳۱	۰/۱۲۹	−۲٦۱/۱۰۳	۰/۷۳۸	−۲۹۷/۹٦۸
٤	۰/۳۹٤	−۱۸۳/۹۰۹	۰/۹٦۱	−۱۸۱/٦۰۲	۳۲	۰/۱۸۳	−۲۱۳/۱۱	۰/۷۸۱	−۲۲٦/٤۵۹
۵	۰/۲٦۷	−۱۸٤/٤۵۰	۰/۹۷٤	−۱٦۹/٤۸۱	۳۳	۰/۰۰۰	−۱۹۹/۵۵۸	۰/۹۱٤	−۱۹۸/٤۳٤
٦	۰/۳۱٦	−۱۹۰/٤٦۸	۰/۹۵۹	−۱۷۳/۲۸۱	۳٤	۰/۰٤۷	−۲۱۲/٤۷۰	۰/٦۹۸	−۲۲۵/۳۲۸
۷	۰/۲٤۵	−۱۸۷/۵۷۳	۰/۹۲۸	−۲۰۵/۲۰٦	۳۵	۰/۰۰۰	−۲۲۹/۰۳۰	۰/۷٤۰	−۲٤۰/٤٦٦
۸	۰/۳۵۸	−۱۹۱/۸٤۳	۰/۹۵۱	−۱۷۷/٤٤۷	۳٦	۰/۰۰۰	−۲۲۱/۹۲٦	۰/۷۷۰	−۲۲۹/۰۷۲
۹	۰/۰۳۵	−۲۰۳/۵۰٦	۰/۸۷٤	−۲۰۷/۵۵٤	۳۷	۰/۱٦۱	−۲۰٤/۱۳۵	۰/۸۵۷	−۲۳۵/٤٦۲
۱۰	۰/۰۰۰	−۲۱۲/۹۹٦	۰/۷۳۸	−۲۳۵/۷۷۵	۳۸	۰/۱۷٦	−۱۸۸/۷۸۵	۰/۸٤۳	−۲۱۲/۷۸٦
۱۱	۰/۳۷۸	−۱۷۳/۹۰۵	۰/۹٦٤	−۱۸۱/۱۲۳	۳۹	۰/۰۰۰	−۲۱۳/۹۸۳	۰/۸۵۵	−۲۰۱/۸۹۵
۱۲	۰/۳۷۸	−۱۹٦/۵۳۱	۰/۹٤۷	−۱٦۹/۳۲۰	٤۰	۰/۰۹۷	−۲۱۱/۰٤۳	۰/۸۷۸	−۲۰۰/٦۸۲
۱۳	۰/۰۰۰	−۲٦٦/۱۲۳	۰/۸۰۰	−۳۰۹/۱۱۸	٤۱	۰/۰۷۰	−۲۷۹/۵۰۱	۰/٦۸۵	−۲٦٤/٦٦۱
۱٤	۰/۳۸۲	−۱۷۲/۱٤۰	۰/۹۵٤	−۱۹۸/۱۰۸	٤۲	۰/۲۳٦	−۲۱۵/٤۹۰	۰/۸۵۲	−۱۹۸/٦۱۳

سؤال	R^2بارعاملی	سهم بارعاملی درتابع زیان	R^2آستانه	سهم آستانه در تابع زیان	سؤال	R^2بارعاملی	سهم بارعاملی در تابع زیان	R^2آستانه	سهم آستانه در تابع زیان
۱۵	۰/٤۰۸	−۱۹۳/٦۹۰	۰/۹۱۵	−۱۹٤/٥۹۷	٤۳	۰/۱۰٥	−۲۱٤/٥۲٦	۰/۸۸٥	−۱۹٦/۰٦۸
۱٦	۰/٤۱۰	−۱۷۷/۸۸٥	۰/۹۷۰	−۱۷٤/٦٤٤	٤٤	۰/۳۸۳	−۱۹۸/۲۷٦	۰/۹۲۱	−۲۲۰/۹۹۹
۱۷	۰/۰۰۰	−۲۰۹/٥٥۱	۰/۸۹۳	−۱۸۳/٥۷۷	٤٥	۰/۰۰۰	−۲۲۲/۸٤۷	۰/۸۸۲	−۱۹۷/۱٥۷
۱۸	۰/۳٥۹	−۱۷۷/۲۲۹	۰/۹٤۲	−۱٦۹/۸٦٦	٤٦	۰/۰۰۰	−۲۱۹/٦٥۷	۰/۷۱۹	−۲۳۷/٤۱٥
۱۹	۰/۱۷۷	−۱۹۸/۳٦۸	۰/۹۰۱	−۱۷۷/٦٦۱	٤۷	۰/۱۳۷	−۲۷٤/٤۸۹	۰/٥۰۲	−۲۳٦/۸٥۳
۲۰	۰/٤۸۳	−۱۷۹/۱۹۰	۰/۹٥۲	−۱۷۹/۰۲۲	٤۸	۰/۱٤٤	−۱۹۸/۲۰۷	۰/۸٤٤	−۲۱۸/٦۱٤
۲۱	۰/۰۰۰	−۱۹٥/۷۲۷	۰/۸۳٤	−۲۰٥/۰٦۱	٤۹	۰/۱۸۱	−۲۱٥/۳۸۹	۰/۸۷۷	−۲۸۱/۹۹۸
۲۲	۰/۰۰۰	−۲۱٥/۷۲٤	۰/۸۱۰	−۲۲٦/٤٦۷	٥۰	۰/۳۲٦	−۲۱٤/۸۷۲	۰/۸۲۰	−۲۸۹/۱۹۹
۲۳	۰/۱۷۲	−۲۰٥/٤۹۰	۰/۸۳۱	−۱۹۰/۷۲٦	٥۱	۰/۳۲۳	−۲۰٦/۸٤۷	۰/۹۲٤	−۱۹۸/۸٦۹
۲٤	۰/۱۱۱	−۲۰٤/۸۷٤	۰/۹۳۷	−۱۹٤/٦۸۸	٥۲	۰/۱٥٥	−۲٤۱/٤٤٤	۰/٦۱۱	−۲۷۹/۰٥۱
۲٥	۰/۱٤۸	−۲۱۷/۹٤۱	۰/۷۹۰	−۲۱٤/٤۳۲	٥۳	۰/۱٥۸	−۲٤۳/۳٥۷	۰/٥٥٥	−۲٦۳/٦٤۷
۲٦	۰/۰۹۳	−۲٥۰/۸٦۳	۰/۳۳٤	−۲۹۷/۲۱۳	٥٤	۰/۱۷۸	−۲۱۰/۱۸٥	۰/۸۰۹	−۲۰۷/٤۸۰
۲۷	۰/۱۹۸	−۱۹٤/۰۳۸	۰/۸٤۳	−۲۱۸/۷۸۸	٥٥	۰/۰۳٥	−۲۰۷/۰۱۲	۰/۸۱٦	−۲۲۹/٦٤۱
۲۸	۰/۰۹۲	−۲۹٥/۱۲٦	۰/٦٤۳	−۲۹۰/٦٥۲					

بحث و نتیجه‌گیری

در این پژوهش، تلاش شد با توجه به نقاط قوت و ضعف و ملاک‌های ارزیابی برازش مدل در آزمون‌های تغییرناپذیری اندازه‌گیری در تعداد زیادی از گروه‌ها در داده‌های دوارزشی، تغییرناپذیری اندازه‌گیری آزمون ریاضی گروه آزمایشی ریاضی‌وفنی در استان‌های ایران آزمون شود. هدف اول پژوهش، بررسی ساختار عاملی آزمون ریاضی گروه آزمایشی ریاضی ۱۳۹٦، نشان داد که ۲٤ سؤال ابتدایی آزمون ریاضی در یک عامل و ۳۱ سؤال انتهایی در عامل دیگر جای گرفتند. بنابراین لازم است بررسی شود که آیا سؤال‌ها در عامل دوم منعکس‌کننده سرعت و محدودیت زمانی در آزمون است. اگر S_W/S_X کوچک‌تر یا مساوی ۰/۱ باشد، آزمون مورد بررسی آزمون سرعت[1] است. در این رابطه w نشان‌دهنده تعداد سؤال‌ها با پاسخ غلط و u نشان‌دهنده تعداد سؤال‌های

[1]. Speed test

انتهایی آزمون، که آزمودنی فرصت نداشته به آنها پاسخ دهد و $x=w+u$ و S_w و S_x انحراف استاندارد w و x است (گالیکسن[1]، ۱۹۵۰). این نسبت در آزمون ریاضی برابر ۰/۳۶۲ که نشان می‌دهد آزمون، آزمون سرعت نیست و بنابراین نمی‌توان گفت سؤال‌های انتهایی آزمون که در عامل دوم قرار می‌گیرند ناشی از سرعت یا محدودیت زمانی است. با توجه به چندبعدی بودن آزمون، به طراحان سؤال پیشنهاد می‌شود، برای سؤال‌ها، جدول محتوایی شناختی طراحی کرده و مهارت‌های مورد نیاز و گام‌های رسیدن به جواب را مشخص کنند تا امکان تحلیل DIF مبتنی بر چند بعد، برای پژوهشگران فراهم شود.

بررسی هدف دوم پژوهش، مقایسه‌پذیری ساختار عاملی سؤال‌های آزمون ریاضی گروه آزمایشی ریاضی ۱۳۹۶ در استان‌های ایران بر مبنای رویکرد تحلیل عاملی نشان داد هر سه مدل شکلی، متریک و اسکالر، برازش مناسبی دارند. آزمون ریاضی دارای تغییرناپذیری اسکالر است بنابراین سازه‌ای یکسان را در استان‌ها اندازه می‌گیرد و میانگین عوامل زیربنایی را می‌توان در استان‌ها مقایسه کرد. اما مشکلی که هنگام استفاده از نمونه‌های بزرگ (بزرگ‌تر از ۵۰۰) به وجود می‌آید این است که تحلیل عاملی چندگروهی همواره مدل اسکالر را نشان می‌دهد (بالن، ۱۹۹۰). بنابراین روش تحلیل عاملی تأییدی چندگروهی برای بررسی تغییرناپذیری در پژوهش‌های مربوط به آزمون‌های ورودی دانشگاه‌ها که تعداد شرکت‌کنندگان در آزمون زیاد است، مناسب نیست. در این موارد از روش بهینه‌سازی ترازبندی می‌شود که مجموع تغییرپذیری اندازه‌گیری در بارهای عاملی و آستانه‌ها در تمام زوج‌های احتمالی از استان‌ها به حداقل مقدار می‌رسد، به‌طوری که تعداد اندکی پارامتر با تغییرپذیری و تعداد زیادی پارامتر با تغییرناپذیری تقریبی وجود داشته باشد.

نتایج حاصل از بررسی هدف سوم، مقایسه میانگین عوامل در میان استان‌ها و رتبه‌بندی و گروه‌بندی استان‌های ایران بر اساس میانگین عوامل زیربنایی، نشان داد به ترتیب استان‌های تهران، اصفهان، خراسان رضوی با تفاوت معنی‌دار نسبت به سایر استان‌ها بیشترین میانگین عاملی را در عامل اول (حساب دیفرانسیل و انتگرال) به دست آورده‌اند و زنجان، گلستان، اردبیل، هرمزگان، لرستان، خراسان شمالی، سیستان‌وبلوچستان، بوشهر، و کهکیلویه‌وبویراحمد در رتبه آخر قرار گرفته‌اند. در عامل

[1]. Gulliksen

دوم نیز به ترتیب استان تهران، اصفهان، و خراسان رضوی در رتبه‌های اول و استان‌های اردبیل، سیستان‌وبلوچستان، و هرمزگان در رتبه‌های آخر قرار دارند.

به‌طور کلی، عملکرد آزمودنی‌ها در استان‌های اصفهان، تهران، خراسان رضوی، فارس، مازندران، یزد و البرز در بیشتر سؤال‌ها نسبت به سایر استان‌ها بهتر است. این استان‌ها، برخوردار و توسعه‌یافته هستند و آزمودنی‌ها در این استان‌ها به‌طور میانگین، معدل کتبی دیپلم بالاتری نسبت به سایر استان‌ها دارند و درصد پاسخ درست به سؤال‌ها برای این آزمودنی‌ها بالاتر است و سؤال‌ها برای آنها اغلب ساده‌تر است. از آنجایی که آزمون ورودی دانشگاه‌ها باید تمامی سطوح توانایی را پوشش دهد نمی‌توان سؤال‌ها یا محتواهایی که در میان این ۷ استان با سایر استان‌ها دارای DIF است حذف کرد. قرار دادن سؤال‌های متفاوت برای مناطق آموزشی مختلف می‌تواند در این زمینه راه‌گشا باشد. از طرفی با اختصاص سهمیه مناطق بومی برای استان‌های محروم باز هم درصد پذیرش در آزمون سراسری در این استان‌ها نسبت به استان‌های برخوردار کمتر است. بنابراین یک پژوهش کیفی و آسیب‌شناسانه می‌تواند در زمینه شناسایی عوامل تاثیرگذار بر کارکرد افتراقی سؤال‌ها در میان استان‌ها یاری‌رسان باشد. به معلمان و داوطلبان آزمون سراسری، در هر استان، پیشنهاد می‌شود در دوره دبیرستان برنامه‌ریزی‌های لازم را درمورد محتوای سؤال‌ها با کارکرد افتراقی اعمال کنند تا از ضعف‌های آموزشی و یادگیری در این مباحث کاسته شود.

در زمینه شناسایی توسعه‌یافتگی و رتبه‌بندی استان‌های کشور، پژوهش‌هایی انجام گرفته است؛ مثلاً میرغفوری و همکاران (۱۳۸۹) نشان دادند استان‌های تهران و یزد و آذربایجان شرقی بهترین وضعیت را در زمینه شاخص‌های کتابخانه‌ای دارند که به‌عنوان شاخصی برای توسعه‌یافتگی در نظرگرفته شده است. صفائی‌پور و مودت (۱۳۹۲) با ارزیابی استان‌های ایران با تأکید برشاخص‌های اجتماعی-اقتصادی، دریافتند که استان‌های تهران و سمنان، به ترتیب بیشترین و کمترین شاخص ترکیبی توسعه انسانی را دارند. پیشنهاد می‌شود شاخص‌های توسعه‌یافتگی و اجتماعی-اقتصادی در سال‌های اخیر در استان‌ها و همچنین ارتباط بین این شاخص‌ها با رتبه‌بندی حاصل از نتایج این پژوهش، بررسی شود.

نتایج حاصل از بررسی هدف چهارم روش چندگروهی بهینه‌سازی ترازبندی در شناسایی سؤال‌های دارای تغییرپذیری اندازه‌گیری، نشان داد که تعداد ۱٤(۲٦٪) سؤال، تغییرناپذیری اندازه‌گیری آستانه در تمامی‌استان‌ها دارند که ۳۷/۵ درصد سؤال‌ها در

عامل اول و۱۶ درصد سؤال‌ها در عامل دوم برای تمامی‌استان‌ها دارای تغییرناپذیری اندازه‌گیری تقریبی آستانه‌ها هستند. تعداد ۲۰ سؤال از ۲٤ سؤال عامل اول (۸۳٪) و تعداد ۲۲ سؤال از ۳۱ سؤال عامل دوم (۷۱٪) دارای تغییرناپذیری اندازه‌گیری بارهای عاملی هستند. به‌طور کلی، ٤۲ سؤال از ۵۵ سؤال (۷٦٪) تغییرناپذیری اندازه‌گیری متریک دارند.

سؤال‌های عامل دوم که سؤال‌هایی با محتوای هندسه و جبرواحتمال و ریاضیات گسسته است، در این آزمون نسبت به عامل اول کارکرد افتراقی بیشتری در میان استان‌ها دارند، بنابراین برای مقایسه آزمودنی‌ها در استان‌ها بهتر است از سؤال‌های عامل اول استفاده شود. پیشنهاد می‌شود نمره‌گذاری به تفکیک ابعاد آزمون‌ها یا با وزن‌دهی متفاوت به سؤال‌ها در هر عامل، باشد تا تأثیر مضرات آزمون برای استان‌های کم‌برخودار کمتر شود. سؤال‌های ۱۰۵(توابع نمایی، نقطه برخورد)، ۱۱٦(تعیین ریشه معادله درجه سوم)، ۱۱۱(حدتوابع)، ۱۰٤(نمودار تابع مثلثاتی) به ترتیب سؤال‌هایی با بالاترین شاخص R^2، یعنی تغییرناپذیرترین سؤال‌ها در پارامتر آستانه هستند و بنابراین سؤال‌های مناسبی برای مقایسه افراد در استان‌ها هستند. نتایج پژوهش نشان داد که میانگین عوامل زیربنایی آزمون ریاضی در استان تهران با سایر استان‌ها دارای تفاوت معنی‌دار است، بنابراین با استفاده از روش‌های DIF در مقایسه‌های دوگروهی کارکرد افتراقی سؤال‌ها در استان تهران به صورت دوبه‌دو با سایر استان‌ها مقایسه شد. سؤال ۱۱٦ در مقایسه جداگانه استان‌ها با استان تهران علیه بیشتر استان‌ها دارای کارکرد افتراقی است در حالی که این سؤال در روش بهینه‌سازی سؤال مناسبی تشخیص داده شده است. کارکرد افتراقی این سؤال در سایر استان‌ها به صورت دوبه‌دو(مانند استان سمنان و گیلان و ...) بررسی شد، که نشان داد در سایر مقایسه‌های زوجی که شامل استان تهران نباشد، DIF وجود ندارد. بنابراین سؤال ۱۱٦ از سؤال‌هایی است که باعث تفاوت معنی‌دار میانگین عاملی استان تهران با سایر استان‌ها شده است. باید توجه کرد که برای بررسی دقیق کارکرد افتراقی سؤال‌ها در تعداد گروه‌های زیاد تنها استفاده از روش ترازبندی کافی نیست، سؤال ۱۱٦ به دلیل اینکه در بیشتر استان‌ها کارکرد افتراقی نداشته است به‌عنوان سؤال تغییرناپذیر در روش بهینه‌سازی شناسایی شده است.

با توجه به اینکه استان محل سکونت آزمودنی، عاملی برای کارکرد افتراقی در سؤال‌ها بود، به منظور سیاست‌گذاری و برنامه‌ریزی در امر آموزش و سنجش و مقایسه عملکرد دانش آموزان در استان‌های مختلف پیشنهاد می‌شود به جای استفاده از نمره

کل آزمون، از سؤال‌ها یا محتواهای دارای تغییرناپذیری اندازه‌گیری در استان‌ها در هر حیطه آموزشی استفاده شود.

پژوهشگران باید توجه داشته باشند که تنها آزمون تغییرناپذیری اندازه‌گیری و شاخص‌های برازش مدل‌ها نمی‌تواند دلیل قطعی برای مقایسه‌پذیری سازه‌ای ساختار عاملی سؤال‌های یک آزمون باشد، به‌ویژه هنگامی که تعداد گروه‌ها و حجم نمونه مورد بررسی زیاد است. در این موارد لازم است از سایر آزمون‌های تغییرناپذیری و کارکرد افتراقی دو گروهی برای گروه‌هایی که گمان می‌رود دارای تفاوت معنی‌دار هستند، استفاده شود.

در تحلیل مجموعه داده‌های پیچیده که شامل افراد از ملت‌ها، ایالت‌ها، اقوام و فرهنگ‌های مختلف هستند، تمرکز مطالعات در پیشینه پژوهش، فراتر از شناسایی DIF است و به سمت توضیح منابع DIF حرکت کرده است (آلبانو و رودریگیز[1]، ۲۰۱۳). توجه نکردن به DIF و منابع آن می‌تواند به استنباط‌های نادرست در مقایسه نمره‌های آزمون در گروه‌های مختلف منجر شود. انجام مطالعه‌ای که هم بر شناسایی منابع DIF و هم بر انتخاب راهبرد مناسب برای بررسی تأثیر منابع شناسایی شده تمرکز داشته باشد، مطالعه‌ای جامع در زمینه کارکرد افتراقی خواهد بود. در نهایت، تغییرناپذیری اندازه‌گیری و سنجش کارکرد افتراقی سؤال در روایی نمره‌های آزمون مسئله‌ای کلیدی است. با توجه به افزایش وابستگی سیاست‌گذاران آموزشی بر ارزیابی‌های بین‌المللی مانند آزمون‌های تیمز و پیرلز و آزمون‌های ملی مثل آزمون‌های استخدامی و آموزش عالی، بی‌توجهی به منابع DIF می‌تواند به استنباط‌هایی اشتباه در مقایسه نمرات آزمون‌ها منجر شود. سیاست‌گذاران باید هنگام تصمیم‌گیری در مورد برنامه درسی، منابع یا آموزش بر مبنای هر مقایسه مستقیم با استفاده از سؤال‌های آزمون‌ها، مراقب باشند، روایی مقایسه بین گروه‌ها همیشه باید پیش از مقایسه نمرات بررسی شود.

---

[1]. Albano & Rodriguez

منابع

ایزانلو، بلال؛ بازرگان، عباس؛ فرزاد، ولی اله؛ صادقی، ناهید؛ کاوسی، امیر (۱۳۹۳). تفکیک ابعاد متعامد از خوشه‌های سؤال بر اساس هشت روش تعیین بعد در داده‌های دوارزشی: مورد آزمون ریاضی رشته ریاضی فیزیک کنکور ۹۱-۹۲. *فصلنامه اندازه‌گیری تربیتی*، ۵(۱۸)، ۲۰۷-۲۴۰.

چگینی، مریم؛ خدایی، ابراهیم؛ فرزاد، ولی‌اله؛ ایزانلو، بلال (۱۳۹۸). داده‌های گمشده در آزمون‌های سراسری ورود به دانشگاه: مبانی نظری و شواهد مبتنی بر داده‌های واقعی. *مطالعات اندازه‌گیری و ارزشیابی آموزشی*، ۹(۲۶)، ۷۱-۱۰۸.

سلیمی، جمال؛ پاسالاری، حامد (۱۳۹۶). نقش ویژگی‌های اجتماعی و تحصیلی دانش‌آموختگان دبیرستانی استان هرمزگان در موفقیت آنان در آزمون ورودی دانشگاه‌ها. *مطالعات اندازه‌گیری و ارزشیابی آموزشی*، ۷(۱۸)، ۹۱-۱۲۵.

شریفی یگانه، نگار (۱۳۹۱). ارزیابی کارکرد افتراقی جنسیتی سؤالات آزمون ریاضی با استفاده از دو روش مانتل-هنزل و نظریه سؤال-پاسخ. *فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی*، ۱(۲)، ۵۳-۷۶.

صفائی‌پور، مسعود؛ مودت، الیاس (۱۳۹۲). ارزیابی استان‌های ایران با تأکید بر شاخص‌های اجتماعی-اقتصادی و شاخص ترکیبی توسعه انسانی با استفاده از تکنیک TOPSIS و GIS . *مطالعات ساختار و کارکرد شهری*، ۳(۱)، ۱۱-۲۷.

گرامی‌پور، مسعود؛ رضایی، احمد؛ رمضان صدر، اعظم؛ نوروزی، لیلا (۱۳۹۵). کنش افتراقی سؤال در آزمون‌های سازمان سنجش آموزش کشور بر حسب ویژگی‌های جمعیت‌شناختی داوطلبان کنکور سراسری. *فصلنامه اندازه‌گیری تربیتی*، ۷(۲۶)، ۱۱۰-۱۲۲.

گرامی‌پور، مسعود؛ رمضانی صدر، اعظم؛ رضایی، احمد؛ نوروزی، لیلا؛ مختاریان، فرانک (۱۳۹۶). کنش افتراقی سؤال (DIF) و آزمون (DTF) مرتبط با جنسیت در آزمون‌های کنکور سراسری سازمان سنجش آموزش کشور. *فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی*، ۷(۱۹)، ۳۵-۶۳.

میرغفوری، سیدحبیب‌الله؛ طحاری مهرجردی، محمدحسین؛ بابایی، حمید (۱۳۸۹). شناسایی وضعیت توسعه‌یافتگی و رتبه‌بندی استان‌های کشور از لحاظ دسترسی به شاخص‌های بخش کتابخانه‌ای. *فصلنامه کتابداری و اطلاع‌رسانی*، ۱۳(۳)، ۲۴۳-۲۷۰.

۸۰ درصد قبولی‌های کنکور اهل کدام استان‌ها هستند؟، (۱۳۹۸، ۲۰ مرداد)، *دنیای اقتصاد*، شماره ۳۵۵۸۹۳۱.

Asparouhov, T., & Muthén, B. O. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling*, 21, 495–508. doi:10.1080/10705511.2014.919210.

Bollen, K. A. (1990). Overall fit in covariance structure models: two types of sample size effects. *Psychol*, *Bull*. *107*(2), 256. Doi: 10.1037/0033-2909.107.2.256.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, 34, 155–175. Doi: 10.1080/10705510701301834

Cheung, G.W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233–255. Doi: 10.1207/S15328007SEM0902_5.

Diana, G., & Tommasi, Ch. (2002). Cross-validation methods in principal component analysis: acomparison. *Statistical Methods & Applications*, *11*, 71-82.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology*, *70*, 662-680.

Flake, J. K., McCoach, D. B. (2017). An investigation of the alignment method with polytomous indicators under conditions of partial measurement invariance. Structural Equation Modeling: A Multidisciplinary Journal, 25, 56–70. doi:10.1080/10705511.2017.1374187

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research*, 23, 267–269.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, 18, 117-144.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary*, *1*(6), 1-55.

Jennrich R. I. (2006). Rotation to simple loadings using component lossfunctions: The oblique case. *Psychometrika*, 71, 173-191.

Kim, E. S., Cao, CH., Wang, Y., & Nguyen, D. T. (2017) Measurement Invariance Testing with Many Groups: A Comparison of Five Approaches. *Structural Equation Modeling: A*

Multidisciplinary Journal, *24*(4), 524-544, DOI: 10.1080/10705511.2017.1304822

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

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, 58, 525–543.

Muthén, B., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiplegroup and growth modeling in Mplus*. Mplus Web Note #4.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide*. Eighth Edition. Los Angeles, CA: Muthén & Muthén

Oliveri, M. E., & Von Davier, M. (2014) Toward Increasing Fairness in Score Scale Calibrations Employed in International Large-Scale Assessments. *International Journal of Testing*, *14*(1), 1-21, DOI: 10.1080/15305058.2013.825265

Revelle, W. (2015). *Psych: Procedures for personality and psychological research* (1.5.8) [Computer software package and manual]. Evanston, IL: Northwestern University. Retrieved from https://cran.r-project.org/web/packages/psych.

Rock, D. A., Werts, C. E., & Flaugher, R. L. (1978). The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations. *Multivariate Behavioral Research*, *13*, 403-418.

Rosseel, Y. (2012). Lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software*, *48*(2), 1–36. http://www.jstatsoft.org/v48/i02/.

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational & Psychological Measurement*, 74, 31–57. Doi: 10.1177/ 0013164413498257.

Stout, W., Froelich, A., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 357-376). New York, NY: Springer-Verlag.

Svetina, D., & Rutkowski, L. (2014). Detecting differential item functioning using generalized logistic regression in the context of large-scale assessments. *Large-scale Assess Education*, *2*(4). https://doi.org/1186/10/s40536-014-0004-5