مجله‌ی مطالعات آموزش و یادگیری
دوره‌ی هشتم، شماره‌ی دوم، پاییز و زمستان ۱۳۹۵، پیاپی ۷۱/۲، صفحه‌های ۹۲- ۱۲۱
(مجله‌ی علوم اجتماعی و انسانی سابق)

بررسی اعتبار نمرات سوابق تحصیلی، نمرات آزمون سراسری و نمرات ترکیبی در پذیرش دانشجویان با استفاده از تئوری تعمیم‌پذیری*

سلیمان ذوالفقارنسب غلامرضا یادگارزاده*** احسان جمالی**** ابراهیم خدایی*****
سازمان سنجش آموزش کشور

چکیده

بنا بر تصویب قانون پذیرش دانشجویان در سال‌های آینده بر اساس افزایش تدریجی سهم سوابق تحصیلی آن‌ها در ترکیب با نمره‌ی آزمون سراسری، مشخص کردن کم و کیف ترکیب این دو نوع نمره برای انتخاب شایسته‌ی متقاضیان ورود به دانشگاه‌ها، یکی از مهم‌ترین چالش‌های تصمیم‌گیرندگان در سازمان سنجش آموزش کشور بوده است. این پژوهش به‌منظور بررسی اعتبار نمرات حاصل از سوابق تحصیلی، آزمون سراسری و ترکیب آن‌ها بر اساس تئوری تعمیم‌پذیری صورت گرفته است. روش تحقیق به‌صورت توصیفی بوده و در آن نمرات سوابق تحصیلی یک سال آخر ۵۶۰۸ دانش‌آموزان گروه ریاضی و فنی به همراه رتبه‌ی دروس مشابه در آزمون سراسری آن‌ها مورد تجزیه‌وتحلیل قرار گرفته است. نتایج نشان داد که هم نمرات سوابق تحصیلی و هم نمرات آزمون سراسری برای رتبه‌بندی و ارتقای افراد به مراحل بعدی آموزش به‌تنهایی دارای ضرایب تعمیم‌پذیری بهینه‌ای هستند؛ *واریانس واقعی* هر سری از این نمرات جداگانه می‌توانند افراد را برای ادامه تحصیل رتبه‌بندی کنند. اما ترکیب این دو نوع نمره اثربخش نیست. آزمون کنکور هنجار مرجع و با چولگی مثبت است و سوابق تحصیلی از آزمون‌های مختلف به‌دست می‌آید که ملاک مرجع و با چولگی منفی هستند. ترکیب آن‌ها ایجاد یک توزیع دونمایی می‌کند که در برآورد نمره‌ی واقعی واریانس صفر یا منفی تولید می‌کند و ضرایب تعمیم‌پذیری به‌دست آمده را بی‌اعتبار می‌سازد. درنهایت تبدیل نمرات به رتبه‌های صدکی برای حذف چولگی و سپس تبدیل رتبه‌ها به توزیع نمرات استاندارد Z برای ترکیب کردن توزیع دو نوع نمره، باعث افزایش واریانس واقعی و بهبود ضرایب تعمیم‌پذیری و اعتمادپذیری می‌شود. همچنین در تهیه‌ی نمره‌ی کل ترکیبی برای این گروه آزمایشی باید تعداد دروس عمومی آن‌ها را کاهش داد و هم‌زمان تک نمره‌ی ریاضی در خرده آزمون‌های اختصاصی آن‌ها را به ۴ لایه نمره تفکیک کرد.

واژه‌های کلیدی: تئوری تعمیم‌پذیری، خطای اندازه‌گیری، نمره‌ی جهانی، واریانس واقعی.

---

* این مقاله خلاصه‌ای است از پروژه‌ای با همین نام که در گروه پژوهشی سنجش و اندازه‌گیری مرکز تحقیقات و ارزشیابی سازمان سنجش آموزش کشور در تابستان ۱۳۹۳ به پایان رسیده است.

** کارشناس پژوهشی سازمان سنجش آموزش کشور، مرکز تحقیقات و ارزشیابی(نویسنده مسؤول) salarnik2001@yahoo.com
*** استادیار سازمان سنجش آموزش کشور
**** استادیار سازمان سنجش آموزش کشور
***** معاون وزیر و ریاست سازمان سنجش آموزش کشور

تاریخ دریافت مقاله: ۹۴/۶/۲۳ تاریخ دریافت مقاله‌ی نهایی: ۹۴/۱۱/۵ تاریخ پذیرش: ۹۵/۲/۲۰

مقدمه

در سال‌های اخیر موضوع کاهش سهم نمره‌ی آزمون سراسری در فرآیند گزینش دانشجو مورد مناقشه‌ی گروه‌های مختلف جامعه ازجمله سیاست‌گذاران عرصه‌ی علمی و سیاسی بوده است. با توجه به تصویب «قانون[۱] سنجش و پذیرش دانشجو در دانشگاه‌ها و مراکز آموزش‌عالی کشور» و الزام وزارت علوم، تحقیقات و فناوری و سازمان سنجش به اعمال تدریجی تأثیر سوابق تحصیلی در پذیرش دانشجویان، در صورت وجود هرگونه ناهماهنگی بین نمرات آزمون‌های ملاک مرجع[۲]، سوابق تحصیلی و نمرات آزمون‌های هنجار مرجع[۳] سراسری می‌تواند مشکلاتی مثل عدم پایایی و اعتمادپذیری را برای ترکیب کردن این نوع نمره‌ها پیش‌بینی کرد.

باید توجه داشت که آزمون‌های ملاک مرجع مثل سوابق تحصیلی عملکرد دانش‌آموزان را با توجه به یک حیطه‌ی خوب تعریف شده می‌سنجند و توصیفی از دانش، مهارت‌ها و یا رفتار یک دانش‌آموز در دامنه‌ای معین و خوب تعریف شده از اهداف و محتوای آموزشی را فراهم می‌کنند. نتایج به‌دست آمده از آزمون‌های ملاک مرجع بر عملکرد دیگر دانش‌آموزان در آزمون وابسته نیست. این در حالی است که در آزمون‌های هنجار مرجع عملکرد یک دانش‌آموز بر اساس عملکرد دیگر دانش‌آموزان تفسیر می‌شود (بوند[۴]، ۱۹۹۶). آزمون‌های هنجار مرجع برای رتبه‌بندی دانش‌آموزان «روی منحنی نرمال» طراحی شده‌اند و نه برای این‌که آیا دانش‌آموزان به استانداردها و یا اهداف آموزشی دست یافته‌اند.

بنابراین آزمون‌های هنجار مرجع را نباید برای ارزیابی رسیدن دانش‌آموزان به استانداردهای آموزشی به‌کار برد (انجمن تحقیقات آموزشی آمریکا[۵]، ۱۹۹۹). به‌هرحال به‌کاربردن هم‌زمان نمرات آزمون‌های ملاک مرجع و هنجار مرجع و ساخت یک نمره‌ی ترکیبی[۶] با پرسش‌هایی همراه است. بر این اساس پژوهش حاضر در چارچوب تئوری تعمیم‌پذیری[۷] به بررسی میزان اعتمادپذیری به ترکیب این دو نوع نمره برای پذیرش افراد پرداخته است. این تئوری که ترکیب کارآمدی از تئوری کلاسیک آزمون‌سازی و تحلیل واریانس می‌باشد، یک مدل اندازه‌گیری است که می‌توان از آن برای بررسی مشارکت نسبی تک‌تک منابع خطای اندازه‌گیری یک نمره‌ی کل ترکیبی مثل نمرات چندگانه حاصل از دروس مختلف، موقعیت‌های مختلف اندازه‌گیری و هر شرایط عملیاتی دیگری که بر اندازه‌گیری‌های رفتاری و روان‌شناختی حاکم باشد به‌علاوه‌ی تعامل بین تک‌تک خطاها را بررسی و نمره‌ی کل بهینه را پیش‌بینی کرد (کرانباخ، گلاسر، نادا و راجارات نان ۱۹۷۲؛ برنان، ۲۰۰۳، ۲۰۰۴ و ۲۰۱۱).

اگر پایایی را صحت، دقت و تکرارپذیری یک اندازه‌گیری در زمینه‌ی پیشرفت تحصیلی در نظر بگیریم، در تئوری کلاسیک آزمون‌سازی روش‌هایی همچون بازآزمایی، فرم‌های موازی، دو نیمه آزمون،

کودر ریچاردسون ۲۰ و کودر ریچاردسون ۲۱ و همچنین آلفایα کرانباخ برای بررسی پایایی مورد استفاده قرار گرفته است (کرانباخ و همکاران، ۱۹۷۲). محدودیت‌های زیادی در استفاده از تئوری کلاسیک آزمون سازی برای برآورد پایایی نمرات ترکیبی وجود دارد. در تئوری کلاسیک آزمون‌سازی واریانس نمرات مشاهده شده x به دو بخش تقسیم می‌شود: واریانس سیستماتیک که از آن به‌عنوان واریانس نمره واقعی τ یاد می‌شود و بخشی که واریانس تصادفی است و واریانس خطا ε نامیده می‌شود. در عمل استفاده از تئوری کلاسیک نشان داده که واریانس خطا ε ساختار یکپارچه‌ای ندارد و عموماً از منابع چندگانه ناشی می‌شود. به همین دلیل، برآورد خطا و برآورد پایایی یعنی نسبت واریانس واقعی به واریانس مشاهده شده می‌تواند بر اساس طرح یا شیوه‌ی جمع‌آوری داده‌ها متفاوت باشد. این محدودیت‌ها باعث شد افرادی مثل کرانباخ، گلاسر، نادا و راجارات نان (۱۹۷۲) راهی برای بررسی همه جانبه‌ی نمره‌هایی که حاصل ترکیب چند شاخص اندازه‌گیری مختلف هستند فراهم سازند. تئوری تعمیم‌پذیری تحت عنوان آزاد کردن تئوری پایایی[۸] از مفروضه‌های محدودکننده برابری (میانگین، واریانس و کواریانس) توانست تئوری کلاسیک آزمون‌سازی را به کمک تحلیل واریانس به اندازه‌گیری‌هایی که آمارهای آن‌ها نابرابر بود گسترش دهد (نابوام، ۱۹۸۴). در رابطه با تحلیل نتایج آزمون‌ها، «تئوری تعمیم‌پذیری تنها مفروضه‌ی آزمون‌های موازی که به‌طور تصادفی از جهانی مشابه نمونه‌گیری شده‌اند را در نظر می‌گیرد» (برنان، ۱۹۹۲ و ۲۰۰۳). سرانجام، درحالی‌که تئوری کلاسیک آزمون‌سازی بر تصمیم‌گیری‌های نسبی یعنی رتبه‌بندی ترتیبی افراد آزمون‌شونده مبتنی است، تئوری تعمیم‌پذیری این برتری را دارد که بین تصمیم‌گیری نسبی (یا هنجار مرجع مثل آزمون کنکور) و تصمیم‌گیری مطلق (یا ملاک مرجع مثل آزمون‌های معلم ساخته در ایران که نمره برش ۱۰ از ۲۰ ملاک است) تمایز قائل می‌شود و برای هر یک به‌طور جداگانه ضرایب تعمیم‌پذیری[۹] $E\hat{\rho}^2_{\delta}$ و شاخص اعتمادپذیری[۱۰] $\hat{\Phi}$ را ارائه می‌دهد (شیولسون و وب، ۲۰۰۶). در تئوری تعمیم‌پذیری مفهوم پایایی در تئوری کلاسیک جای خود را به مفهومی گسترده‌تر و انعطاف‌پذیرتری به‌نام تعمیم‌پذیری می‌دهد. به جای این‌که بپرسیم با چه اطمینانی نمره‌ی مشاهده شده با نمره‌ی واقعی هماهنگ است (پایایی در تئوری کلاسیک)؟ در تئوری تعمیم‌پذیری می‌پرسیم با چه اطمینانی می‌توان نمره‌ی مشاهده شده را به رفتار فرد در جهانی از موقعیت‌های تعریف شده تعمیم داد (اعتبار در تئوری تعمیم‌پذیری)؟ بنابراین سؤال مربوط به پایاییِ تکرار یک نمره در تئوری کلاسیک، به سؤال مربوط به دقت تعمیم[۱۱] و یا تعمیم‌پذیری یک نمره منتج می‌شود (کرونباخ و همکاران، ۱۹۷۲).

در این تئوری به پایایی به‌عنوان حد و اندازه‌ای نگریسته می‌شود که از یک مشاهده به جهانی از

مشاهدات می‌توان تعمیم داد (کرانباخ، گلاسر و همکاران، ۱۹۷۲). ازاین‌رو در تئوری تعمیم‌پذیری، نمره‌ی مشاهده شده را به‌عنوان نمره‌ی جهانی[۱۲] می‌خوانند (شی ولسون و وب، ۱۹۸۱). مفید بودن و اثربخشی یک نمره یا همان واحد جهانی به‌طور کل وابسته است به حد و اندازه‌ای که به ما اجازه می‌دهد به‌طور معتبری، به رفتار فرد در گستره‌ای از موقعیت ها–در این تحقیق آزمون‌ها منظور است– تعمیم دهیم (کرونباخ و همکاران، ۱۹۷۲). این فرآیند دربرگیرنده‌ی استفاده از یک چهارچوب مفهومی برای برآورد منابع خطای اندازه‌گیری در متن اندازه‌گیری است که در آن برای جداسازی و برآورد مؤلفه‌های واریانس مرتبط با آن‌چه که سطوح یا بندهای[۱۳] اندازه‌گیری و لایه‌ها یا شرایط[۱۴] آن نامیده می‌شود از تحلیل واریانس (ANOVA) استفاده می‌شود؛ تقریباً می‌توان گفت که در یک مطالعه تعمیم-پذیری G مؤلفه‌های واریانس برآورد شده برای نمره $\sigma^2_{X_{pi}}$ هر فرد را می‌توان به این صورت تفسیر کرد که چقدر افراد بر اساس شرایط حاکم بر موقعیت اندازه‌گیری (مثل تعداد سؤالات i، خرده آزمون‌های t، نوع تکلیف، متن، موقعیت آزمون، نمره‌گذاران و نظایر این‌ها) که می‌توان آن‌ها را سطح و یا بند نامید با یکدیگر متفاوتند. هرکدام از این بندها یا سطوح می‌تواند یک منبع واریانس یا پراش σ^2_i باشد که رتبه‌بندی واقعی حاصل از تفاوت‌های سیستماتیک توانایی افراد σ^2_p را مخدوش می‌سازد (براون، ۲۰۰۵). به‌عبارتی میانگین مربعات به‌دست‌آمده از ANOVA برای برآورد مؤلفه‌های واریانس حاصل از سطوح و لایه‌هایی که در هر سطح آشیان گرفته‌اند به‌کار می‌رود، که در مباحث مربوط به پایایی در تئوری کلاسیک هرگز به این جامعیت و با چنین جزییاتی قابل بررسی نبوده است (شیولسون و وب، ۱۹۸۱). در این مسیر شیولسون و وب (۱۹۹۱) روش‌های تئوری تعمیم‌پذیری را به‌صورت نظری مورد بحث قرار دادند و برنان (۱۹۸۳) روش‌های انجام تئوری تعمیم‌پذیری را تا حدود بسیار زیادی به‌صورت عملیاتی گسترش داد.

به‌هرحال چهار سؤال اساسی در این پژوهش که در چهارچوب تئوری تعمیم‌پذیری به آن‌ها پاسخ داده شده، به ترتیب عبارتند از این‌که (۱) آیا می‌توان تنها نمرات دوران دبیرستان را به‌عنوان یک شاخص معتبر برای ورود افراد به دانشگاه در نظر گرفت؟ (۲) آیا می‌توان نمرات کنکور را به‌تنهایی به‌عنوان یک شاخص معتبر برای ورود افراد در نظر گرفت؟ (۳) ترکیب نمرات سوابق تحصیلی با نمرات کنکور می‌تواند باعث افزایش واریانس واقعی و کاهش واریانس خطا شود؟ (۴) افزایش یا کاهش تعداد آزمون‌های مختلف چه تأثیری بر اعتبار اندازه‌گیری می‌گذارد؟

روش پژوهش

طرح این تحقیق توصیفی می‌باشد. هدف این نوع طرح‌های تحقیقاتی توصیف پدیده‌های مورد بررسی است. اجرای تحقیق توصیفی می‌تواند صرفاً برای شناخت بیشتر شرایط موجود و کمک کردن به فرآیند تصمیم‌گیری باشد (سرمد، بازرگان و حجازی، ۱۳۸۳). در این تحقیق نمرات آزمون سراسری و سوابق تحصیلی و ترکیب دو نوع نمره مورد بررسی قرار گرفته است.

شرکت‌کنندگان پژوهش

گروه نمونه و جامعه‌ی این تحقیق یکی می‌باشد و به‌عبارتی همه کسانی هستند که در آزمون گروه ریاضی و فنی در جلسه‌ی آزمون سراسری ۱۳۹۲ به خرده آزمون‌های کنکور جواب داده‌اند و برای آن‌ها کارنامه صادر شده و تعداد آن‌ها برابر با ۵۶۰۸ داوطلب بوده است.

روش اجرا

برای بررسی اعتبار این نمرات و ترکیب آن‌ها از روش تعمیم‌پذیری تک‌متغیره با الگوی $p \times t$ بر اساس نمودار (۱) و روش چندمتغیره با الگوی $p \times (t: r)$ بر اساس نمودار (۳) که در آن لایه‌های آزمون‌های t که فرض می‌شود از جهانی ممکن از آزمون‌های موازی به‌صورت تصادفی انتخاب شده‌اند، در سطوح r کلی‌تر دروس عمومی r_1 و اختصاصی r_2 هم‌آشیان شده‌اند، استفاده شده است. لازم به ذکر است که برای هماهنگی با پیشینه‌ی پژوهشی تئوری تعمیم‌پذیری از i به‌عنوان شاخص سؤال استفاده شده است. اما درواقع واحد تحلیل در این تحقیق نمره‌ی آزمون t بوده است نه سؤال i.

علایم عبارتند از:

p = نشانگر هرفرد،

t = نمره هر خرده آزمون و نمره معادل با آن درس در سوابق تحصیلی –نه نمره‌ی سؤالات i

r = سطوح آزمون عمومی و اختصاصی

در تئوری تعمیم‌پذیری هر سطح نسبت به لایه‌های پایین‌تر خود دارای اثرات تثبیت‌شده هستند و لایه‌های آن دارای اثرات تصادفی. به‌عنوان مثال در تحلیل‌های مبتنی بر الگوی $p \times t$ آزمون‌شوندگان p اثر تثبیت‌شده[۱۵] هستند و آزمون‌های t اثرات تصادفی[۱۶]. به همین ترتیب در تحلیل‌های مبتنی بر طرح‌های هم‌آشیان $p \times (t: r)$، آزمون‌دهندگان p برای دروس عمومی r_1 و اختصاصی r_2 اثر تثبیت‌شده هستند و دروس عمومی r_1 و اختصاصی r_2 دارای اثرات تصادفی برای آزمون‌دهندگان p و در یک مرحله پایین‌تر دروس عمومی r_1 و اختصاصی r_2 برای آزمون‌های t آشیان‌گرفته در آن‌ها اثر تثبیت‌شده هستند و لایه‌های آزمون‌های t دارای اثرات تصادفی هستند. به‌عبارتی‌دیگر در الگوی $p \times (t: r)$ هر

سطح نسبت به لایه‌های خود دارای اثرات تثبیت‌شده و لایه‌ها نسبت به سطوح دارای اثرات تصادفی می‌باشند.

سطوح تصادفی دربرگیرنده‌ی «مجموعه‌ای از شرایط» است که از مجموعه‌ای بی‌نهایت از شرایط مربوط به یک بند یا سطح به‌صورت تصادفی نمونه‌گیری شده‌اند و فرض بر این است که با هر نمونه‌ی دیگری از شرایط، قابل تعویض هستند. به‌عنوان مثال فرض بر این است که هر آزمون t در مجموعه خود r نماینده‌ای است از جهانی از سؤالات ممکن در آن درس که به‌صورت تصادفی انتخاب شده‌اند (برنان، ۲۰۰۱، ۲۰۰۳). لایه‌های آزمون‌های t و سطوح دروس عمومی r_1 و دروس اختصاصی r_2 در جدول (۱) آمده‌اند. همچنین داده‌ها با نرم افزار mGENOV نسخه‌ی ۲/۱ (برنان، ۲۰۰۱) تحلیل شده‌اند.

جدول ۱- سطوح و لایه‌های آزمون‌ها

	سطوح آزمون (r)	
	عمومی r1	اختصاصی r2
لایه‌های آزمون (t)	۱-زبان فارسی ۲-ادبیات فارسی *(۳)-آزمون ادبیات (۴)*	۱-جبر و احتمال ۲-هندسه ۳-حسابان *(۴)-آزمون ریاضی (۴)*
	۱-عربی (۲)-آزمون عربی (۲)	۱-فیزیک و آزمایشگاه *(۲)-آزمون فیزیک (۳)*
	۱-تعلیمات دینی و قرآن (۲)-آزمون معارف (۳)	۱-شیمی و آزمایشگاه *(۲)-آزمون شیمی (۲)*
	۱-زبان خارجی *(۲)-آزمون زبان (۲)*	

همچنین برای شناسایی وزن واقعی و مؤثر هر آزمون، از نمرات بدون ضریب اسمی در تحلیل‌ها استفاده شده است. ضرایب اسمی همان وزن‌هایی هستند که به‌عنوان ضریب به دروس اصلی هر رشته می‌دهند تا وزن آن‌ها را در مجموعه‌ی آزمون بالاتر ببرند. به این دلیل از نمرات بدون ضریب اسمی در تحلیل استفاده شده تا بتوان وزن مؤثر و واقعی نمره‌ی هر آزمون را بدون این‌که با ضرایب اسمی مخدوش شوند آشکار کرد (ذوالفقارنسب، خدایی و یادگارزاده، ۱۳۹۲). در جدول (۱) ضریب اسمی هر درس در آزمون سراسری در پرانتز آمده است. مواد آزمون‌هایی که به‌صورت برجسته و کج هستند مربوط به آزمون سراسری با ضرایب اسمی آن است. به‌علاوه چون داده‌های مورد نیاز در دسترس

نبوده‌اند لایه‌های هر درس از دو نمره بیشتر تشکیل نشده‌اند به‌استثنای دروس فارسی (۳ درس) و دروس ریاضی (۴ درس).

شیوه‌ی تحلیل داده‌ها

دو رویکرد عمده در این تئوری برای بررسی متغیرها و یا مؤلفه‌های یک اندازه‌گیری وجود دارد:

الف) تئوری تعمیم‌پذیری تک‌متغیره

ب) تئوری تعمیم‌پذیری چندمتغیره

تئوری تعمیم‌پذیری تک‌متغیره برای تحلیل منابع چندگانه واریانس خطا و قابلیت اعتماد به یک نمره که می‌تواند حتی نمره یک سؤال i و یا نمره یک آزمون t باشد به کار می‌رود. به‌عبارت‌دیگر در روش تک‌متغیره، هدف این است که قابلیت اعتماد یا تعمیم‌پذیری نمرات حاصل از یک اندازه‌گیری و حدومرزی که در آن تحت شرایط یکسان یک نمره تکرارپذیر باشد مورد بررسی قرار می‌گیرد (برنان، ۲۰۰۳). این نمره می‌تواند نمره‌ی یک سؤال i یا نمره‌ی یک خرده آزمون t با n سؤال i باشد؛ در مقایسه با تئوری کلاسیک ما پایایی تک‌تک سؤال‌ها i و پایایی کل آزمون t را داریم.

در یک طرح تعمیم‌پذیری کاملاً متقاطع فرد × سؤال که در آن همه‌ی افراد p به همه‌ی سؤالات i یک آزمون t پاسخ می‌دهند و تنها سؤالات سطوح یا بندها هستند (نمودار ۱) نمره X_{pi} مشاهده شده یک فرد p روی سؤال‌ها i را می‌توان به مؤلفه‌های زیر تجزیه کرد (شیولسون و وب، ۲۰۰۵):

$$X_{pi} = \mu + (\mu_p - \mu) + (\mu_i - \mu) + (X_{pi} - \mu_p - \mu_i + \mu) \qquad (۱)$$

μ = میانگین بزرگ[۱۷] روی هر دو جامعه افراد و سؤالات جهانی[۱۸] می‌باشد و برای تمام افراد یک مقدار یکسان است.

μ_p = میانگین افراد که می‌توان آن را مقدار مورد انتظار $E(X)$ نمره‌ی مشاهده شده‌ی یک فرد p روی سؤالات جهانی i تعریف کرد و در تئوری تعمیم‌پذیری به‌عنوان نمره‌ی جهانی مشخص می‌شود و μ_i = میانگین جمعیت روی سؤال i می‌باشد (برنان، ۲۰۰۳).

به‌جز میانگین بزرگ μ، بخش‌ها یا مؤلفه‌های تشکیل‌دهنده‌ی هر نمره، دارای یک توزیع مربوط به خود می‌باشند. به‌این‌ترتیب که توزیع نمره‌ی افراد ($\mu_p - \mu$) میانگینی برابر با صفر و واریانسی برابر با $\sigma_p^2 = E_p(\mu_p - \mu)^2$ دارد که واریانس نمره‌ی جهانی خوانده می‌شود. به‌طور مشابه، مؤلفه‌ی مربوط به سؤال $(\mu_i - \mu)$ میانگینی برابر با صفر و واریانسی برابر با $\sigma_i^2 = E_i(\mu_i - \mu)^2$ دارد. مؤلفه‌ی مربوط به باقیمانده‌ها $(X_{pi} - \mu_p - \mu_i + \mu)$ میانگینی برابر با صفر و واریانسی برابر با

$\sigma^2_{p_i,e} = E_i E_p (X_{pi} - \mu_p - \mu_i + \mu)^2$ دارد که بیانگر تعامل خطای نامنظم یا برآورد نشده[۱۹] فرد × سؤال است. پس «واریانس کل» مجموعه‌ای از نمرات مشاهده شده $\sigma^2_{X_{pi}} = E_p E_i (X_{pi} - \mu)^2$ را می‌توان به‌صورت فرمول (۲) تجزیه کرد:

$$\sigma^2_{X_{pi}} = \sigma^2_p + \sigma^2_i + \sigma^2_{p_i,e} \qquad (۲)$$

σ^2_p واریانس نمره‌ی جهانی $\sigma^2_p = E_p(\mu_p - \mu)^2$

σ^2_i واریانس نمره‌ی سؤال $\sigma^2_i = E_i(\mu_i - \mu)^2$

$\sigma^2_{p_i,e}$ واریانس باقیمانده $\sigma^2_{p_i,e} = E_i E_p (X_{pi} - \mu_p - \mu_i + \mu)^2$

مؤلفه‌ی واریانس باقیمانده $\sigma^2_{p_i,e}$ منعکس کننده‌ی اثر تعامل فرد × سؤال و دیگرخطاهای تصادفیِ تعریف نشده است که می‌توان با استفاده از تحلیل واریانس مبتنی بر نمرات مشاهده شده در یک طرح، فرد × سؤال که در آن نمونه‌ای تصادفی از n_p آزمون شونده به یک آزمون مشتمل بر n_i سؤال که به‌طور تصادفی انتخاب شده‌اند، آن را برآورد کرد. در مقایسه با ضریب پایایی $\rho_{xx'}$ در تئوری کلاسیک، ضریب تعمیم‌پذیری $E\rho^2$ را می‌توان به‌صورت فرمول (۳) نوشت (وب، شیولسون و هارتل ۲۰۰۶؛ هی، ۲۰۰۹؛ برنان، ۲۰۰۱، ۲۰۰۳):

$$E\rho^2 = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_\delta} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{pI,e}} = \frac{\sigma^2_p}{\sigma^2_p + \frac{\sigma^2_{pi,e}}{n_i'}} \qquad (۳)$$

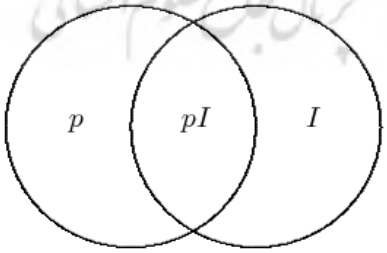
σ^2_p = واریانس نمره‌ی جهانی

σ^2_δ = واریانس خطای نسبی δ که مبتنی است بر نسبت واریانس خطا $\sigma^2_{pi,e}$ به تعداد n_i' مشاهدات جهانی یا: $\sigma^2_{pI,e}$ که در آن I شاخص میانگین سؤالات i است.

آزمون‌گیرندگان ممکن است آزمون اجراشده را به‌صورت هنجار مرجع و یا ملاک مرجع به‌کار بگیرند. یا بخواهند نمرات ترکیبی تهیه کنند که از آزمون‌های ملاک مرجع کلاسی و هنجار مرجع مثل آزمون سراسری تشکیل شده باشد. همان‌طور که پیشتر بیان شد در آزمون‌های هنجار مرجع رتبه‌بندی افراد نسبت به یکدیگر مطرح است و به دنبال آن دقت این رتبه‌بندی برای محقق بسیار اهمیت دارد. در آزمون‌های ملاک مرجع میزان چیرگی افراد بر محتوای موضوعی معین اهمیت دارد. این‌که چه تعداد

افراد قبول می‌شوند مهم نیست. افراد از یک نمره برش معین به بالا تفکیک می‌شوند. به‌عنوان مثال در آزمون راهنمایی و رانندگی میزان چیرگی بر موضوعات رانندگی اهمیت دارند و نمره برش ۱۷ نمره قبولی است. در تئوری تعمیم‌پذیری این امکان هست که هم اعتبار رتبه‌بندی افراد نسبت به یکدیگر را بر مبنای آزمون‌های هنجار مرجع مشخص کرد و هم دقت و اعتبار یک نمره در یک نقطه برش معین برای تفکیک افراد به قبول یا رد را بر مبنای آزمون‌های ملاک مرجع مشخص کرد. بنابراین بر اساس نمرات به‌دست‌آمده محقق می‌تواند هم‌زمان دو نوع تصمیم بگیرد: تصمیم‌گیری نسبی (هنجار مرجع) و تصمیم‌گیری مطلق (ملاک مرجع[۲۰]). یک تصمیم‌گیری نسبی یا هنجار مرجع مبتنی است بر رتبه‌بندی ترتیبی[۲۱] افراد که اعتبار آن با ضریب $E\rho^2$ مشخص می‌شود و یک تصمیم‌گیری مطلق یا ملاک مرجع مبتنی است بر یک مقدار عملکرد یا نمره برش معین در توزیع نمرات که با ضریب $\hat{\Phi}$ مشخص می‌شود. بنابراین تئوری تعمیم‌پذیری بین دو نوع ضریب تمایز قائل می‌شود: یک ضریب تعمیم‌پذیری G[۲۲] برای تصمیم‌گیری‌های نسبی $E\rho^2$ و یک شاخص اعتمادپذیری phi[۲۳] برای تصمیم‌گیری‌های مطلق $\hat{\Phi}$.

به‌علاوه در تئوری تعمیم‌پذیری این امکان هست که بر اساس طرح‌های تصمیم‌گیری D[۲۴] مشخص کرد که برای رسیدن به یک «دقت اندازه‌گیری» معین چند سؤال به آزمون اضافه شود و یا در مجموعه آزمون‌ها مثل کنکور سطوح r و لایه‌های آزمون‌های t به چه صورت باشند تا یک اندازه‌گیری با بیشترین دقت و کمترین خطا به‌دست آید. برای کاهش واریانس خطا و افزایش اعتبار نمرات می‌توان به همان سبکی که در فرمول اسپیرمن براون برای افزایش پایایی در تئوری کلاسیک آزمون‌سازی بیان شده است تعداد سطوح یا لایه‌ها را افزایش داد. در نمودار (۱) فضای یک طرح ساده‌سازی شده فرد × سؤال یعنی $p \times i$ برای یک خرده آزمون n سوالی i و تعامل واریانس p و iها را می‌توان دید.



نمودار ۱- ون برگرفته از رابرت آل. برنان

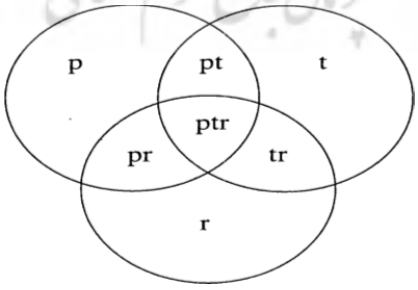
p= واریانس سیستماتیک افراد بر اساس توانایی آن‌ها،

I= واریانس سؤالات

pI= تعامل خطای فرد و سؤالات

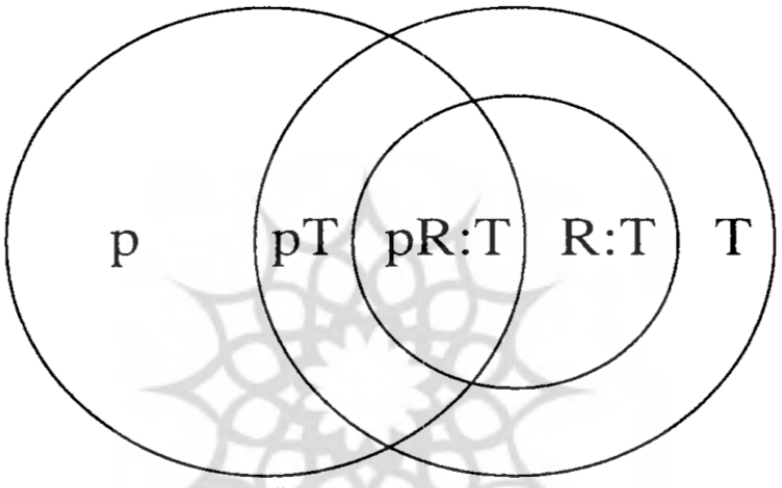
در مقابل تئوری تعمیم‌پذیری چند متغیره برای بررسی هم‌زمان ویژگی‌های روان‌سنجی و اعتبار اندازه‌گیری‌های چندگانه مثل نمرات جهانی به‌دست‌آمده از خرده آزمون‌های مختلف توسعه داده شده‌اند. در این موقعیت‌ها نمرات جهانی[۲۵] چندگانه که هرکدام حاصل سازه‌های مختلف عملیاتی شده به‌وسیله‌ی آزمون‌های t مختلف هستند یک نمره‌ی کل ترکیبیِ جامع یا جهانی را تشکیل می‌دهند که سهم هر آزمون t در تشکیل آن نمره می‌تواند دارای خطاهای متعددی باشد (شیولسون و وب، ۱۹۸۱؛ برنان، ۲۰۰۱؛ وب، شیولسون و هرتل، ۲۰۰۶). تئوری تعمیم‌پذیری چند متغیره‌ی واریانس و کوواریانس مشاهده‌شده حاصل از این نمرات جهانی را به مؤلفه‌های آن تفکیک می‌کند و منابع خطاهای چندگانه را بدین‌صورت بررسی می‌کند. باید اضافه کرد که برخلاف نمره‌ی کل در تئوری کلاسیک که بر اندازه‌ی نمره کلِ[۲۶] X حاصل از سرجمع نمرات چند سؤال i یا چند آزمون t وابسته است به‌عنوان مثال $\sum_{t=1,t\neq i}^{n} X_t$ ، تئوری تعمیم‌پذیری بر اندازه‌ی میانگین $\overline{X}$ نمرات[۲۷] چند سؤال i یا چند آزمون t مبتنی است به‌عنوان مثال $\frac{1}{n}\sum_{t=1,t\neq i}^{n} X_t$ (وب، شیولسون و هارتل، ۲۰۰۶).

در نمودار (۲) فضای یک طرح چند متغیره $p \times t \times r$ با دو سطح یا دو بند متقاطع $t \times r$ و تعامل واریانس‌های هرکدام را می‌توان دید. در این طرح فرض بر این است که چند آزمون t به‌وسیله‌ی چندین داور r نمره‌گذاری شده‌اند. یعنی همه‌ی نمره‌گذاران یا داوران همه‌ی آزمون‌ها را تصحیح و نمره‌گذاری می‌کنند. طبیعتاً هم سطح آزمون‌های t با چند لایه و هم سطح نمره‌گذاران r با چند لایه منابع خطا هستند.



نمودار ۲- ون برگرفته از رابرت آل. برنان

در تحقیق حاضر با توجه به شیوه‌ی دسته‌بندی آزمون‌ها و شیوه‌ی نمره‌گذاری و ساخت نمره‌ی کل طبیعتاً نمره‌گذاری وجود ندارد. صرفاً نمرات آزمون‌های کنکور با سوابق تحصیلی مشابه آن‌ها در لایه‌های مربوط به خود-یعنی یک درس معین- جمع می‌شوند و درنهایت در دو سطح نمره‌ی عمومی و اختصاصی نمره‌ی جهانی ساخته می‌شود. بنابراین سطح مربوط به آزمون‌های t در دل سطح کلی‌تر R یعنی دروس عمومی و اختصاصی آشیان می‌گیرند. به همین دلیل با توجه به شیوه‌ی طرح تصمیم‌گیری نمره‌ی جهانی به‌صورت طرح آشیان‌دار $p \times (t: r)$ و با توجه به نمودار (۳) تهیه می‌شود.



نمودار ۳- ون برگرفته از رابرت آل. برنان

همان‌طور که پیشتر بیان شد، در تئوری تعمیم‌پذیری دو نوع مطالعه یا بررسی صورت می‌گیرد: الف) مطالعه‌ی تعمیم‌پذیری G برای برآورد مؤلفه‌های واریانس یک مشاهده‌ی معتبرِ جهانی که توضیح داده شد و ب) مطالعه‌ی تصمیم‌گیری D. در مرحله‌ی بررسی‌های تصمیم‌گیری پس از این‌که یک نمره‌ی جهانی برای تعمیم مشخص شد، می‌توان بر اساس آن نمره، طرح‌های مختلف تصمیم‌گیری طراحی کرد مثل آشیان‌دار کردن سطوح و یا افزایش و کاهش لایه‌های درون هر آشیان و بدین ترتیب برآوردهایی از «میانگینِ مؤلفه‌های واریانس» صورت می‌گیرد و بر اساس این طرح‌های مختلف تصمیم‌گیری D میزان اعتمادپذیری به نمرات بر اساس ضرایب اعتماد phi محاسبه می‌شود. به‌طور خلاصه استفاده از چند مطالعه‌ی D و مشاهده‌ی نتایج حاصل از آن‌ها کمک می‌کند، تا محقق مشخص کند چند سؤال در یک آزمون و یا چند خرده آزمون در مجموعه‌ی آزمون نیاز است تا ضریب اعتمادِ موردنیاز برای یک اندازه‌گیری کم خطا در یک برنامه‌ی آزمون‌گیری به‌دست آید. همچنین در مطالعات تصمیم‌گیری D محقق می‌تواند سطح‌ها یا بندهایی را کاهش و یا افزایش دهد یا طرح‌های ضربدری

(×) یا متقاطع را به طرح‌های آشیان‌دار (:) تبدیل کند تا به مقدار بهینه‌ای از اعتماد به یک نمره‌ی جهانی برسد. چون فرض بر این است که سؤالات یا خرده آزمون‌ها از جهانی از سؤالات و یا خرده آزمون‌های موازی تهیه شده‌اند می‌توان از روی محاسبه‌ی واریانس نمرات موجود بدون هدر دادن هزینه و زمان برای اجرای آزمون‌های متعدد و یا تعداد سؤالات بیشتر اما با اثربخشی کمتر میزان تعمیم‌پذیری نمرات را پیش‌بینی کرد. همچنین برای برنامه‌های سنجشی بعدی تصمیم‌گیری‌های بهینه‌ای در رابطه با تعداد سؤالات یک آزمون و یا تعداد خرده آزمون‌های لازم برای ساخت یک نمره‌ی کل ترکیبی بهینه انجام داد. در نمودار (۳) یک نمای ساده از این نوع تصمیم‌گیری نمایش داده شده است که در آن برای ساخت دو نوع نمره‌ی آزمون عمومی r_1 و اختصاصی r_2، لایه‌های دروس t در یک سطح کلی‌تر آشیان‌دار می‌شوند؛ به‌عنوان مثال آزمون‌های چندگانه‌ی ادبیات T در دل آزمون‌های عمومی R آشیانه دارند.

نتایج پژوهش

برای بررسی اولیه‌ی هر دسته آزمون (کنکور و سوابق تحصیلی) به تفکیک دروس عمومی و اختصاصی تحلیل‌های جداگانه‌ای از مجموعه نمرات آن‌ها انجام شده است تا کمیت و کیفیت اولیه‌ی نمرات سوابق تحصیلی و آزمون کنکور به ترتیب مشخص شود. به‌طور کلی درصد ٪ واریانس p و ضرایب تعمیم‌پذیری $E\hat{p}^2_{\delta}$ نمرات هر مجموعه آزمونِ سوابق تحصیلی (جدول ۲) و هر مجموعه آزمونِ سراسری (جدول ۳) به‌طور مجزا در سطح بهینه‌ای قرار داشته‌اند.

برای نمرات ترکیبی سوابق و کنکور دو گونه تحلیل صورت گرفته است. ابتدا طرح تک‌متغیره که هر درس با دو لایه (آزمون) و یا بیشتر مشخص می‌شدند (جدول ۴ و ۵) و سپس طرح‌های چندمتغیره‌ی «هم‌آشیان» که در آن هر درس با لایه‌های (یا آزمون‌های) خود در مجموعه‌ی کلی دروس عمومی r_1 و یا اختصاصی r_2 هم‌آشیان شده‌اند (جدول ۶، ۷ و ۸). در ادامه نتایج هر بررسی آمده است.

منابع واریانس سوابق تحصیلی: بر اساس جدول (۲) سه منبع واریانس برای هر یک از دروس عمومی و اختصاصی سوابق تحصیلی برآورد شده است. این دو درس که هر سطح آن دارای پنج لایه و یا درس متفاوت است، بر اساس تئوری تعمیم‌پذیری تک‌متغیره تحلیل شده‌اند و پارامترهای نمره جهانی (ترکیبی) آن در ستون آخر آمده است. همان‌طور که می‌بینیم واریانس نمره‌ی جهانی دروس عمومی ۶۸٪ و برای دروس اختصاصی ۷۳٪ است و واریانس آزمون‌ها (لایه‌ها) ۵٪ و ۴٪ و اثرات دیگر (یا واریانس خطا) برای این دو به ترتیب ۲۷٪ و ۲۳٪ می‌باشد.

جدول ۲- منابع واریانس «سوابق تحصیلی» با دو سطح r تثبیت‌شده که هر یک دربرگیرنده‌ی نمره‌ی ۵ درس عمومی و ۵ درس اختصاصی است و بر اساس الگوی t × p تحلیل شده‌اند (مقادیر گرد شده‌اند).

	(G-study)	دروس عمومی (۵ آزمون)			دروس اختصاصی (۵ آزمون)			ترکیبی
منبع واریانس	Source of variation	N	واریانس	٪	N	واریانس	٪	
دانش‌آموزان	٪person(p)	۵۶۰۸	۸۱۹۱۱/۱۷	(٪۶۸)	۵۶۰۸	۱۳۲۳۴۰/۵۵	(٪۷۳)	۱۰۱۴۱۳/۵۲
آزمون‌ها	٪test(t)	۵	۵۸۶۱/۶۵	(٪۵)	۵	۶۴۰۱/۲۴	(٪۴)	==
اثرات دیگر یا خطا: تعامل فرد-آزمون و دیگر خطاهای تصادفی	٪pt,e	==	۳۲۳۱۶/۵۳	(٪۲۷)	==	۴۱۴۱۲/۲۸	(٪۲۳)	==
Coefficients	ضرایب			(٪۱۰۰)			(٪۱۰۰)	
ضریب تعمیم‌پذیری G (تصمیم‌گیری نسبی)	$E\hat{p}^2_\delta$	۰/۹۲۶			۰/۹۴۱			۰/۹۶۵
شاخص اعتمادپذیری (تصمیم‌گیری مطلق)	$\hat{\Phi}$	۰/۹۱۴			۰/۹۳۲			۰/۹۶۴
سهم درصدی از واریانس کل ٪۱۰۰	٪	٪۴۳/۷۸			٪۵۶/۲۲			٪۱۰۰
وزن	w-weights	۰/۵۰			۰/۵۰			۱

منابع واریانس آزمون سراسری: بر اساس جدول (۳) سه منبع واریانس برای هر یک از دروس عمومی و اختصاصی آزمون سراسری برآورد شده است. این دو درس که سطح عمومی آن دارای چهار لایه و سطح اختصاصی آن دارای سه لایه یا درس متفاوت است، بر اساس تئوری تعمیم‌پذیری تک-متغیره تحلیل شده‌اند. همان‌طور که می‌بینیم واریانس نمره‌ی جهانی p دروس عمومی ٪۵۴ و برای دروس اختصاصی ٪۷۳ است و واریانس آزمون‌ها (لایه‌ها) ٪۹ و ٪۱ و اثرات دیگر (یا واریانس خطا) برای این دو به ترتیب ٪۳۷ و ٪۲۵ می‌باشد.

جدول ۳- منابع واریانس «آزمون سراسری» با دو سطح r تثبیت‌شده که هر سطح دربرگیرنده‌ی ۴ آزمون عمومی و ۳ آزمون اختصاصی است و بر اساس الگوی p × t تحلیل شده‌اند.

	(G-study)	آزمون عمومی (۴)			اختصاصی (۳)			ترکیبی
منبع واریانس	Source of variation	N	واریانس	٪	N	واریانس	٪	
دانش‌آموزان	٪person(p)	۵۶۰۸	۲۴۹/۲۳	(٪۵۴)	۵۶۰۸	۱۲۹/۲۸	(٪۷۳)	۱۸/۵۳
آزمون‌ها	٪test(t)	۴	۳۹/۶۷	(٪۹)	۳	۲/۱۲	(٪۱)	==
اثرات دیگر یا خطا: تعامل فرد-آزمون و دیگر خطاهای تصادفی	٪pt,e	==	۱۷۰/۵۸	(٪۳۷)	==	۴۴/۸۳	(٪۲۵)	==
Coefficients	ضرایب			(٪۱۰۰)			(٪۱۰۰)	
ضریب تعمیم‌پذیری G (تصمیم‌گیری نسبی)	$E\hat{p}^2_{\delta}$		۰/۸۵۴			۰/۸۹۶		۰/۹۱۷
شاخص اعتمادپذیری (تصمیم‌گیری مطلق)	$\hat{\Phi}$		۰/۸۲۶			۰/۸۹۲		۰/۹۰۳
سهم درصدی از واریانس کل ۱۰۰٪	٪		٪۶۵/۴۵			٪۳۴/۵۵		٪۱۰۰
وزن	w-weights		۰/۵۷			۰/۴۳		۱

منابع واریانس نمرات آزمون‌های ترکیب‌شده: بر اساس جدول (۴) سه منبع واریانس برای هر یک از نمرات دروس سوابق تحصیلی ترکیب‌شده با دروس همسان آن‌ها در کنکور برآورد شده است. در این طرح با ۷ سطح تثبیت‌شده سر و کار داریم و درون هر سطح لایه‌های تصادفی قرار گرفته‌اند به‌عبارتی بر اساس این الگو دروس به دو بخش عمومی و اختصاصی تفکیک نشده‌اند بلکه به‌صورت ۷ درس مجزا (با سطوح تثبیت‌شده) تحلیل شده‌اند. به‌عنوان مثال درس فارسی ۳ لایه دارد (۱-زبان فارسی،۲-ادبیات فارسی و (۳)-آزمون ادبیات).

درس ریاضی ۴ لایه دارد (۱-جبر و احتمال، ۲-هندسه، ۳-حسابان و (۴)-آزمون ریاضی). این کار برای ارزیابی اولیه‌ی ماهیت نمره‌ی هر درس سوابق تحصیلی در ترکیب با درس همسان آن در کنکور انجام شده است. نمرات براساس تئوری تعمیم‌پذیری تک‌متغیره p × t تحلیل شده‌اند و نمره‌ی جهانی ترکیبی آن در ستون آخر آمده است. همان‌طور که می‌بینیم واریانس نمره‌ی جهانی برای دروس فارسی ۴٪ و برای دروس ریاضی ۱۳٪ است و برای دیگر دروس«صفر»می‌باشد. به‌عنوان مثال درس عربی را درنظر بگیریم که از دو نمره‌ی سوابق تحصیلی و نمره‌ی آزمون سراسری تشکیل شده، نشان می‌دهد که ۹۴ درصد تغییرات نمرات افراد ناشی از ترکیب این دو نوع آزمون (ملاک مرجع و هنجار مرجع) است و ۴ درصد ناشی از خطای محاسبه نشده است و واریانس واقعی p که این درس برای تفکیک افراد تولید می‌کند صفر است. این نشان می‌دهد که ترکیب این دو نوع آزمون (هنجار مرجع و ملاک مرجع) و ساخت نمره‌ی کل برای هر درس معتبر نیست. عملاً این دو نوع آزمون در ترکیب باهم تفاوت‌هایی

در نمرات افراد ایجاد می‌کنند که کلِ آن برآیند و یا نتیجه‌ی تفاوت ماهیت ملاک مرجع و هنجار مرجع بودن این دو نوع آزمون است نه واریانس سیستماتیک و مطلوب نمره‌ی جهانی افراد p. این در حالی است که هرکدام از این دو نوع آزمون (سوابق تحصیلی و کنکور) به‌تنهایی در مجموعه‌ی خود دارای اعتبار است و به خوبی افراد را از یکدیگر تفکیک می‌کنند (به جدول ۲ و ۳ نگاه کنید).

ضریب تعمیم‌پذیری $E\hat{\rho}^2_{\delta}$ و شاخص اعتمادپذیری $\hat{\Phi}$ هر درس پایین‌تر از مقدار مورد انتظار است به‌عنوان مثال برای دروس فارسی ضریب تعمیم‌پذیری $E\hat{\rho}^2_{\delta}$=۰/۶۸ و برای شیمی برابر با $E\hat{\rho}^2_{\delta}$=۰/۰۸ است. همچنین شاخص اعتمادپذیری برای دروس فارسی $\hat{\Phi}$ =۰/۱۰ و برای دروس شیمی $\hat{\Phi}$ =۰ است. این دو ضریب برای نمره‌ی جهانی حاصل از ترکیب هفت نمره به ترتیب $E\hat{\rho}^2_{\delta}$=۰/۹۰۷ و $\hat{\Phi}$ =۰/۴۰۳ می‌باشد (به ردیف مربوط به این دو ضریب در جدول (۴) نگاه کنید).

سهم درصدی واریانس دروس از واریانس کل نمره‌ی جهانی به ترتیب برای دروس فارسی ۱۷/۹۰٪ و برای دروس ریاضی ۳۴/۹۹٪ است. مابقی سهم درصدی هر درس از واریانس کل نمره‌ی جهانی در ردیف مربوط به آن آمده است.

جدول ۴- منابع واریانس«دروس سوابق تحصیلی و آزمون سراسری ترکیب‌شده» با ۷ سطح r تثبیت‌شده که هر سطح دربرگیرنده‌ی چند لایه‌ی متفاوت است و بر اساس الگوی p × t تحلیل شده‌اند.

(G-study)	سوابق تحصیلی در ترکیب با نمرات همسان در آزمون سراسری																			
Source of variation	N	۱-زبان فارسی ۲-ادبیات فارسی (۳)-آزمون ادبیات	٪	N	۱-عربی (۲)-آزمون عربی	٪	N	۱-دین و زندگی (۲)-آزمون معارف	٪	N	۱-زبان خارجی (۲)-آزمون زبان	٪	N	۱-جبر و احتمال ۲-هندسه ۳-حسابان (۴)-آزمون ریاضی	٪	N	۱-فیزیک (۲)-آزمون فیزیک	٪	N	۱-شیمی (۲)-آزمون شیمی	٪	ترکیبی
٪person(p)	۵۶۰۸	۳۰۰۳۶/۲۶	(٪۴)	۵۶۰۸	۴۳۶۳/۴۷	(٪۰)	۵۶۰۸	۴۱۱۱/۱۴	(٪۰)	۵۶۰۸	۵۰۳۶/۲۸	(٪۰)	۵۶۰۸	۷۲۸۷۸/۲۹	(٪۱۳)	۵۶۰۸	۳۷۳۰/۰۳	(٪۰)	۵۶۰۸	۲۸۱۵/۶۵	(٪۰)	۳۴۶۲۸/۱۳
٪test(t)	۳	۶۹۲۴۴۳/۷۳	(٪۹۱)	۲	۹۷۷۱۸۵/۷۰	(٪۹۴)	۲	۱۲۵۶۷۰۸/۱۶	(٪۹۶)	۲	۱۰۹۸۳۷۰/۲۹	(٪۹۴)	۴	۴۲۹۶۳۸/۶۸	(٪۷۵)	۲	۷۰۴۸۱۷/۳۲	(٪۸۹)	۲	۹۴۰۱۶۷/۷۵	(٪۸۹)	==
٪pt,e	==	۴۱۹۵۳/۲۴	(٪۵)	==	۵۸۱۳۲/۸۸	(٪۶)	==	۴۳۵۹۴/۵۲	(٪۳)	==	۶۳۱۸۶/۲۸	(٪۵)	==	۶۷۷۸۱/۳۴	(٪۱۲)	==	۸۳۱۴۵/۶۸	(٪۱۱)	==	۶۳۶۳۲/۵۶	(٪۱۱)	==
Coefficients			(٪۱۰۰)			(٪۱۰۰)			(٪۱۰۰)			(٪۱۰۰)			(٪۱۰۰)					(٪۱۰۰)	
$E\hat{\rho}^2_\delta$	۰/۶۸			۰/۱۳			۰/۱۵			۰/۱۳			۰/۸۱			۰/۰۸۲		۰/۰۸			۰/۹۰۷
$\hat{\Phi}$	۰/۱۰			۰/۰۰			۰/۰۰			۰/۰۰			۰/۳۶			۰/۰۰		۰/۰۰			۰/۴۰۳
٪	٪۱۷/۹۰			٪۹/۶۵			٪۷/۵۴			٪۹/۲۴			٪۳۴/۹۹			٪۱۰/۹۰		٪۹/۷۷			۱۰۰
w-weights	۰/۱۷			۰/۱۲			۰/۱۲			۰/۱۲			۰/۲۳			۰/۱۲		۰/۱۲			۱

منابع واریانس نمرات آزمون‌های ترکیب‌شده بر اساس تبدیل نمرات خام به نمرات استاندارد z در ادامه‌ی نتایج تحلیل مبتنی بر ترکیب سوابق تحصیلی و آزمون سراسری به شیوه‌ای آمده است که در آن برای افزایش واریانس p و کاهش واریانس‌های نامطلوبِ دیگر، هم نمرات کنکور پس از کسر ضریب منفی از آن و هم نمرات سوابق تحصیلی، تبدیل به نمرات z استاندارد شده‌اند. اثر مثبت این تبدیل مقیاس نمرات را می‌توان در افزایش بی‌چون‌وچرای واریانس p مشاهده کرد. بر اساس جدول (۵) واریانس نمره‌ی جهانی برای دروس فارسی ۶۶٪ و برای دروس ریاضی ۶۷٪ است. این نشان می‌دهد که تبدیل این دو نوع نمره، به نمرات استاندارد z و ساخت نمره‌ی کل برای هر درس می‌تواند بسیاری از مشکلات ترکیب این دو نوع نمره را حل کند اما نه همه‌ی مشکلات.

ضریب تعمیم‌پذیری برای دروس فارسی ۰/۸۵ =$E\hat{\rho}^2_\delta$ و برای شیمی برابر با ۰/۷۲=$E\hat{\rho}^2_\delta$ است. همچنین شاخص اعتمادپذیری برای دروس فارسی ۰/۸۵= $\hat{\Phi}$ و برای دروس شیمی ۰/۷۲= $\hat{\Phi}$ است. این دو ضریب برای نمره‌ی جهانی حاصل از ترکیب هفت نمره، به ترتیب ۰/۹۷=$E\hat{\rho}^2_\delta$ و ۰/۹۷= $\hat{\Phi}$ می‌باشد (به ردیف مربوط به این دو ضریب در جدول (۵) نگاه کنید).

سهم درصدی واریانس دروس از واریانس کل نمره‌ی جهانی به ترتیب برای دروس فارسی ۱۷/۹۰٪ و برای دروس ریاضی ۲۳/۹۸٪ است. مابقی سهم درصدی هر درس از واریانس کل نمره‌ی جهانی در ردیف مربوط به آن آمده است.

جدول ۵- منابع واریانس «دروس سوابق تحصیلی و آزمون سراسری ترکیب‌شده» با ۷ سطح r تثبیت‌شده که هر سطح دربرگیرنده‌ی چند لایه‌ی متفاوت است و بر اساس تبدیل نمرات به z استاندارد و الگوی p × t تحلیل شده‌اند.

(G-study)	سوابق تحصیلی در ترکیب با نمرات همسان در آزمون سراسری بر اساس تبدیل نمرات به Z استاندارد																			
Source of variation	N	٪	۱-زبان فارسی ۲-ادبیات فارسی (۲)-آزمون ادبیات	N	٪	۱-عربی (۲)-آزمون عربی	N	٪	۱-دین و زندگی (۲)-آزمون معارف	N	٪	۱-زبان خارجی (۲)-آزمون زبان	N	٪	۱-حجر و احتمال ۲-هندسه ۳-حسابان (۴)-آزمون ریاضی	N	٪	۱-فیزیک (۲)-آزمون فیزیک	N	٪	۱-شیمی (۲)-آزمون شیمی	٪	ترکیبی
٪person(p)	۵۶۰۸	۰/۴۶		۵۶۰۸	(۴۶)	۰/۴۶	۵۶۰۸	(۴۶)	۰/۵۶	۵۶۰۸	(۵۶)	۰/۶۲	۵۶۰۸	(۶۲)	۰/۶۷	۵۶۰۷	(۶۷)	۰/۶۵	۵۶۰۷	(۶۵)	۰/۵۶	(۵۶)	۰/۶۲
٪test(t)	۳	۰/۰۰۰۱		۲		۰/۰۰۰۱	۲		۰/۰۰۰۱	۲		۰/۰۱۰۰	۴		۰/۰۰۰۱	۲		۰/۰۰۰۱	۲		۰/۰۰۰۱		
٪pt,e	==	۰/۳۳	(۳۳٪)	==	(۳۳٪)	۰/۳۳	==	(۳۳٪)	۰/۴۳	==	(۴۳٪)	۰/۳۷	==	(۳۷٪)	۰/۳۳	==	(۳۳٪)	۰/۳۴	==	(۳۴٪)	۰/۴۳	(۴۳٪)	
Coefficients			(۱۰۰٪)			(۱۰۰٪)			(۱۰۰٪)			(۱۰۰٪)					(۱۰۰٪)			(۱۰۰٪)		(۱۰۰٪)	
$E\hat{\rho}^2_\delta$	۰/۸۵			۰/۸۰			۰/۷۲			۰/۷۷			۰/۸۹			۰/۷۹			۰/۷۲			۰/۹۷
$\hat{\Phi}$	۰/۸۵			۰/۸۰			۰/۷۲			۰/۷۷			۰/۸۹			۰/۷۹			۰/۷۲			۰/۹۷
٪	٪۱۷/۹۰			٪۱۲/۲۵			٪۱۱/۲۴			٪۱۰/۷۷			٪۲۳/۹۸			٪۱۲/۱۵			٪۱۱/۷۲			۱۰۰
w-weights	۰/۱۷			۰/۱۲			۰/۱۲			۰/۱۲			۰/۲۳			۰/۱۲			۰/۱۲			۱

منابع واریانس نمرات آزمون‌های ترکیب‌شده با دو سطح عمومی و اختصاصی (طرح هم‌آشیان نابرابر): الگویی که در ادامه می‌آید بر اساس طرح تعمیم‌پذیری هم‌آشیان صورت گرفته است که در آن همه‌ی دروس عمومی r_1 و اختصاصی r_2 دو سطح را تشکیل می‌دهند که لایه‌های t این سطوح دربرگیرنده‌ی مؤلفه‌ها یا آزمون‌های دروس عموی و اختصاصی است به‌نحوی‌که هرکدام از درس‌ها از لایه‌های آزمون‌های مختلف سوابق تحصیلی و آزمون کنکور، اما همسان از لحاظ موضوعی تشکیل شده است. در قیاس با پایایی در تئوری کلاسیک آزمون‌سازی فرض بر این است، هنگامی‌که تعداد سؤالات خرده آزمون‌های یک آزمون ترکیبی افزایش می‌یابد و یا تعداد خرده آزمون‌ها در یک مجموعه‌ی آزمون افزایش می‌یابد، خطا کاهش پیدا می‌کند. به همین دلیل کل آزمون‌ها یا لایه‌ها در دو سطح کلی‌تر عمومی و اختصاصی آشیان گرفته‌اند تا دو نمره‌ی عمومی و اختصاصی حاصل شود که هرکدام از آن‌ها دربرگیرنده‌ی لایه‌های مختص به خود است؛ به‌عبارتی لایه‌های هر سطح افزایش یافته است. بر اساس جدول (۶) پنج منبع واریانس برای هر یک از دروس عمومی و اختصاصی آزمون سراسری برآورد شده است. این دو درس که سطح دروس عمومی آن دارای چهار لایه و سطح اختصاصی آن دارای سه لایه یا درس متفاوت است، بر اساس تئوری تعمیم‌پذیری چندمتغیره تحلیل شده‌اند و واریانس نمره‌ی جهانی (ترکیبی) آن در ستون آخر آمده است. همان‌طور که می‌بینیم واریانس نمره‌ی جهانی p روی دروس عمومی ۲٪ و روی دروس اختصاصی ۶٪ است. همچنین آزمون عمومی ۲۸٪ و آزمون اختصاصی ۲۱٪ واریانس تولید کرده‌اند. از طرف دیگر لایه‌های هر آزمون بیشترین واریانس (نامطلوب) را در دل مجموعه‌ی خود (عمومی و اختصاصی) ایجاد کرده‌اند به‌عنوان مثال دروس فارسی، عربی، دینی و زبان انگلیسی در ترکیب باهم ۶۵٪ و دروس ریاضی، فیزیک و شیمی در ترکیب باهم ۶۴٪ به‌علاوه تعامل واریانس افراد با سطوح آزمون عمومی ۱٪ و با سطوح آزمون اختصاصی ۲٪ خطا ایجاد کرده‌اند. اثرات دیگر و یا خطاهای دیگری که عملیاتی نشده‌اند و به‌عبارتی خطای باقیمانده هستند برای دروس عمومی ۳٪ و برای دروس اختصاصی ۸٪ می‌باشد.

ضریب تعمیم‌پذیری $E\hat{\rho}^2_{\delta}$ به تفکیک برای دروس عمومی ۰/۹۲۸ و برای دروس اختصاصی ۰/۹۴۱ است و شاخص اعتمادپذیری $\hat{\Phi}$ برای دروس عمومی ۰/۹۶۴ و برای دروس اختصاصی ۰/۸۹۱ است. این دو ضریب برای نمره‌ی جهانی (حاصل از ترکیب دو نمره‌ی عمومی و اختصاصی) به ترتیب $E\hat{\rho}^2_{\delta}$ =۰/۹۶۵ و $\hat{\Phi}$ =۰/۹۵۵ است (به ستون آخر جدول (۶) نگاه کنید).

سهم درصدی واریانس دروس عمومی و اختصاصی آزمون سراسری از واریانسِ کل نمره‌ی جهانی به ترتیب برای دروس عمومی ۴۴/۳۷٪ و برای دروس اختصاصی ۵۵/۶۳٪ است.

جدول ۶- منابع واریانس «نمره‌ی سوابق تحصیلی و آزمون سراسری ترکیب‌شده» با دو سطح r تثبیت‌شده نابرابر که دربرگیرنده‌ی ۴ آزمون عمومی با لایه‌های (۳ ۲ ۲ ۲)=t آن و ۳ آزمون اختصاصی با لایه‌های (۴ ۲ ۲)=t آن بر اساس الگوی P × (t: r) تحلیل شده‌اند.

	(G-study)	عمومی ۴ سطح (۹ لایه یا آزمون)			اختصاصی ۳ سطح (۸ لایه یا آزمون)			
منبع واریانس	Source of variation	N	واریانس	٪	N	واریانس	٪	ترکیبی
دانش‌آموزان	person(p)	۸۵۶۰	۲۷۱۷۷/۳۸	(٪۲)	۸۵۶۰	۵۳۱۲۸/۵۹	(٪۶)	۳۶۸۴۰/۲۷
سطوح آزمون عمومی و اختصاصی	r	۴	-۴۰۸۵۳۱/۲۴	(٪۲۸)	۳	-۱۸۷۱۶۶/۴۴	(٪۲۱)	==
لایه‌های هر آزمون در سطوح آزمون عمومی و اختصاصی	t:r	۴	۹۴۳۴۳۰/۳۲	(٪۶۵)	۳	۵۸۶۷۸۰/۲۳	(٪۶۴)	==
تعامل فرد و سطوح آزمون عمومی و اختصاصی	pr	==	-۱۳۲۹۵/۰۰	(٪۱)	==	-۱۴۴۹۲/۲۴	(٪۲)	==
اثرات دیگر یا خطا: تعامل فرد-لایه آزمون با سطوح آزمون عمومی و اختصاصی و دیگر خطاهای تصادفی	pt:r,e	==	۴۹۷۶۴/۰۴	(٪۳)	==	۷۰۰۲۴/۴۵	(٪۸)	==
ضرایب	Coefficients			(٪۱۰۰)			(٪۱۰۰)	
ضریب تعمیم‌پذیری G (تصمیم‌گیری نسبی)	$E\hat{p}^2_\delta$	۰/۹۲۸			۰/۹۴۱			۰/۹۶۵
شاخص اعتماد پذیری (تصمیم‌گیری مطلق)	$\hat{\Phi}$	۰/۹۶۴			۰/۸۹۱			۰/۹۵۵
سهم درصدی از واریانس کل ٪۱۰۰	٪	٪۴۴/۳۷			٪۵۵/۶۳			٪۱۰۰
وزن	w-weights	۰/۵۳			۰/۴۷			۱

منابع واریانس نمرات آزمون‌های ترکیب‌شده با دو سطح عمومی و اختصاصی (طرح هم‌آشیان برابر) مبتنی بر مطالعه‌ی تصمیم‌گیری D: در این تحقیق بر اساس نمرات خام و تبدیل نشده به توزیع z، چندین بررسی تصمیم‌گیری و یا به‌عبارتی مطالعه‌ی D صورت گرفته که به علت پرخطا بودن آن‌ها تنها نتایج دو بررسی گزارش می‌شود که در یکی از (منطقی‌ترین) آن‌ها کمترین میزان خطا به‌دست آمده و با افزایش واریانس واقعی p هم‌زمان ضریب $E\hat{p}^2_\delta$ آن برابر با ۰/۹۰ بوده است. یکی از این روش‌های پرخطا (بر اساس جدول ۷) برابر کردن سطوح دروس عمومی و اختصاصی به ۳ سطح می‌باشد: یعنی ۳ به ۳. به‌عبارتی حذف یکی از دروس عمومی به‌طور کل از مجموعه‌ی آزمون کنکور و سوابق تحصیلی. به‌علاوه برابر کردن لایه‌های درونی هر سطح که در آن دروس ریاضی چهارلایه‌ای به ۲ لایه و دروس فارسی سه‌لایه‌ای (در صورت قرار گرفتن در آزمون و حذف نشدن) هم به ۲ لایه کاهش پیدا کند. این مطالعه‌ی D نشان داد که حذف یکی از دروس عمومی از ۴ سطح به ۳ سطح هیچ تأثیری بر واریانس نمره‌ی جهانی ندارد، اما کاهش لایه‌های دروس اختصاصی یعنی کاهش لایه‌های دروس ریاضی از ۴ لایه به ۲ لایه آسیب جدی به واریانس نمره‌ی جهانی وارد می‌کند. این مشکل هنگامی بیشتر می‌شود که بدانیم نمرات دروس سه‌گانه‌ی

ریاضی یعنی جبر و احتمال، هندسه و حسابان روی‌هم ریخته می‌شوند و یک (لایه) نمره با ضریب ۴ در کنکور به خود اختصاص می‌دهد (به جدول (۷) نگاه کنید).

جدول ۷- منابع واریانس «نمره‌ی سوابق تحصیلی و آزمون سراسری ترکیب‌شده» طرح D بالانس با برابر قرار دادن سطوح دروس عمومی و اختصاصی ۳ به ۳ و لایه‌های درون هرکدام (۲ ۲ ۲)=T؛ با کاهش یک لایه از درس عمومی فارسی (۲ ۲ ۲)=T و ۲ لایه از درس اختصاصی ریاضی (۲ ۲ ۲)=T

	(D-study)	عمومی ۳ سطح (۶ لایه یا آزمون)			اختصاصی ۳ سطح (۶ لایه یا آزمون)			
منبع واریانس	Source of variation	N	واریانس	٪	N	واریانس	٪	ترکیبی
دانش‌آموزان	person(p)	۸۵۶۰	۲۷۱۷۷/۳۸	(٪۱۴)	۸۵۶۰	۵۳۱۲۸/۵۹	(٪۳۳)	۳۷۵۹۴/۶۸
سطوح آزمون عمومی و اختصاصی	R	۳	۰		۳	۰		==
لایه‌های هر آزمون در سطوح آزمون عمومی و اختصاصی	T:R	۳	۱۵۷۲۳۸/۳۸	(٪۸۲)	۳	۹۷۷۹۶/۷۰	(٪۶۰)	==
تعامل فرد و سطوح آزمون عمومی و اختصاصی	PR	==	۰		==	۰		==
اثرات دیگر یا خطا: تعامل فرد-لایه آزمون با سطوح آزمون عمومی و اختصاصی و دیگر خطاهای تصادفی	PT:R,E	==	۸۲۹۴/۰۰	(٪۴)	==	۱۱۶۷۰/۷۴	(٪۷)	==
ضرایب	Coefficients			(٪۱۰۰)			(٪۱۰۰)	
ضریب تعمیم‌پذیری G (تصمیم‌گیری نسبی)	$E\hat{p}^2_\delta$		۰/۷۶۶			۰/۸۱۹		۰/۸۸۲
شاخص اعتماد پذیری (تصمیم‌گیری مطلق)	$\hat{\Phi}$		۰/۱۴			۰/۳۲۶		۰/۳۵۳
سهم درصدی از واریانس کل ٪۱۰۰	٪		٪۴۱/۳۷			٪۵۸/۶۳		٪۱۰۰
وزن	w-weights		۰/۵۰			۰/۵۰		۱

منابع واریانس نمرات آزمون‌های ترکیب‌شده با دو سطح عمومی و اختصاصی (طرح هم‌آشیان با لایه‌های نابرابر) مبتنی بر مطالعه‌ی تصمیم‌گیری D: در ادامه‌ی نتایج بررسی D گزارش می‌شود که در آن یک درس عمومی کاملاً حذف شده و سطوح باقیمانده‌ی آن‌ها برای درس فارسی به ۲ لایه‌ی مساوی کاهش پیدا کرده. به‌علاوه درس ریاضی بر اساس مؤلفه‌های آن افزایش یافته و به ۴ بخش یا مؤلفه‌ی ۱-جبر و احتمال، ۲-هندسه، ۳-حسابان، و ۴-آزمون ریاضی تفکیک شده و برای هر مؤلفه نمره‌ای جدا در نظر گرفته شده است. همان‌طور که پیشتر گفته شد نمره‌ی دروس ریاضی مثل جبر، هندسه و حسابان با یک نمره گزارش می‌شود. در جدول (۸) نتایج این بررسی D آمده است. همان‌طور که می‌بینیم واریانس نمره‌ی جهانی p روی دروس عمومی ۱۴٪ و روی دروس اختصاصی ۳۹٪ است. همچنین واریانس آزمون عمومی برابر ۰٪ و آزمون اختصاصی برابر با ۰٪ قرار داده شده است. از طرف دیگر لایه‌های هر آزمون بیشترین واریانس نامطلوب را در دل مجموعه‌ی خود (عمومی و اختصاصی) ایجاد کرده‌اند، به‌عنوان مثال دروس فارسی، عربی، دینی و زبان انگلیسی ۸۲٪ و دروس ریاضی، فیزیک و شیمی ۵۴٪. به‌علاوه تعامل واریانس افراد با سطوح آزمون عمومی برابر با ۰٪ و با سطوح آزمون اختصاصی برابر با ۰٪ قرار داده شده‌اند. اثرات دیگر و

یا خطاهای دیگری که عملیاتی نشده‌اند و به‌عبارتی خطای باقیمانده هستند، برای دروس عمومی ۴٪ و برای دروس اختصاصی ۶٪ می‌باشد. ضریب تعمیم‌پذیری $E\hat{\rho}^2_\delta$ به تفکیک برای دروس عمومی ۰/۷۶۶ و برای دروس اختصاصی ۰/۸۵۸ است و شاخص اعتمادپذیری $\hat{\Phi}$ برای دروس عمومی ۰/۱۴ و برای دروس اختصاصی ۰/۳۹ است. این مقدار برای آزمون‌های سرنوشت‌ساز بسیار پایین است و گزینش افراد بر اساس آزمون‌هایی با ضریب اعتمادپذیری کم درست نیست. این دو ضریب برای نمره‌ی جهانی حاصل از ترکیب دو نمره‌ی عمومی و اختصاصی به ترتیب $E\hat{\rho}^2_\delta$ =۰/۹۰ و $\hat{\Phi}$ =۰/۴۰ است (به ستون آخر جدول (۸) نگاه کنید).

سهم درصدی واریانس دروس عمومی و اختصاصی آزمون سراسری از واریانسِ کل نمره‌ی جهانی به ترتیب برای دروس عمومی ۳۴/۳۶٪ و برای دروس اختصاصی ۶۵/۶۴٪ است.

جدول ۸- منابع واریانس «نمره‌ی سوابق تحصیلی و آزمون سراسری ترکیب‌شده» طرح D بالانس با برابر قرار دادن سطوح دروس عمومی و اختصاصی ۳ به ۳؛ سه آزمون عمومی با لایه‌های (۲ ۲ ۲)=T و افزایش لایه‌های درس ریاضی در مجموعه‌ی دروس اختصاصی به ۴ لایه (۲ ۲ ۴)=T

	(D-study)	عمومی ۳ سطح (۶ لایه یا آزمون)			اختصاصی ۳ سطح (۸ لایه یا آزمون)			
منبع واریانس	Source of variation	N		٪	N		٪	ترکیبی
دانش‌آموزان	person(p)	۵۶۰۸	۲۷۱۷۷/۳۸	(14٪)	۵۶۰۸	۵۳۱۲۸/۵۹	(٪۳۹)	۳۹۵۰۰/۵۵
سطوح آزمون عمومی و اختصاصی	R	۳	۰		۳	۰		==
لایه‌های هر آزمون در سطوح آزمون عمومی و اختصاصی	T:R	۳	۱۵۷۲۳۸/۳۸	(٪۸۲)	۳	۷۳۳۴۷/۵۲	(٪۵۴)	==
تعامل فرد و سطوح آزمون عمومی و اختصاصی	PR	==	۰		==	۰		==
اثرات دیگر یا خطا: تعامل فرد-لایه آزمون با سطوح آزمون عمومی و اختصاصی و دیگر خطاهای تصادفی	PT:R,E	==	۸۲۹۴/۰۰	(٪۴)	==	۸۷۵۳/۰۵	(٪۶)	==
ضرایب	Coefficients			(٪۱۰۰)			(٪۱۰۰)	
ضریب تعمیم‌پذیری G (تصمیم‌گیری نسبی)	$E\hat{\rho}^2_\delta$	۰/۷۶۶			۰/۸۵۸			۰/۹۰
شاخص اعتماد پذیری (تصمیم‌گیری مطلق)	$\hat{\Phi}$	۰/۱۴			۰/۳۹			۰/۴۰
سهم درصدی از واریانس کل ۱۰۰٪	٪	٪۳۴/۳۶			٪۶۵/۶۴			٪۱۰۰
وزن	w-weights	۰/۴۲			۰/۵۷			۱

بحث و نتیجه‌گیری

در این پژوهش مجموعه‌ای از نمرات حاصل از سوابق تحصیلی، آزمون سراسری و ترکیب آن‌ها بر اساس تئوری تعمیم‌پذیری تحلیل شدند. هدف این تحلیل‌ها نخست برآورد مؤلفه‌های واریانس هر آزمون و یا مجموعه‌ی آن‌ها بوده است. این کار در سه بخش الف) سوابق تحصیلی، ب) آزمون سراسری و ج) نمره‌ی ترکیبی حاصل از سوابق تحصیلی و آزمون سراسری صورت گرفت. نتایج اولیه نشان داد، برای ساخت نمره‌ی کل با داشتن ۴ سطح درس عمومی و ۳ سطح

دروس اختصاصی نمی‌توان نمره‌ی بهینه‌ای برای گروه تخصصی ریاضی و فنی تهیه کرد. نمره‌ی کل تهیه شده برای رشته‌های تخصصی نیازمند هم تعداد نمرات دروس تخصصی بیشتر نسبت به دروس عمومی و هم اختصاص وزن بیشتر برای نمرات درس‌های تخصصی هستند. برای رسیدن به یک نمره‌ی کل بهینه- با وجود ثابت بودن شرایط فعلی-یک طرح تصمیم‌گیری ارائه شده که بر اساس آن برای این‌که واریانس واقعی نمرات افزایش پیدا کند، باید سطوح آزمون عمومی از ۴ لایه نمره به ۳ لایه نمره کاهش پیدا کند و نمره‌ی دروس ریاضی از ۱ نمره به ۴ لایه نمره افزایش یابد. حتی با وجود این تغییرات باید تأکید کرد که بدون تبدیل توزیع نمرات این آزمون‌های هنجار مرجع (سراسری) و ملاک مرجع (سوابق تحصیلی) ترکیب نمرات آن‌ها کار خردمندانه‌ای نیست؛ تبدیل نمرات هر درس به رتبه‌های صدکی و سپس تبدیل این رتبه‌ها به توزیع نمرات استاندارد z می‌تواند بسیاری از مشکلات ترکیب این دو نوع نمره را حل کند. در ادامه به پاسخ‌گویی به سؤال‌های پژوهشی این تحقیق پرداخته شده است.

با توجه به سؤال اول این پژوهش مبنی بر این‌که «۱- اعتبار نمرات مربوط به سوابق تحصیلی با توجه به میزان خطای آزمون‌های (عمومی و اختصاصی) r مختلف، چقدر است؟ G- study»

نتایج حاصل از تحلیل داده‌های سوابق تحصیلی گروه آزمایشی فنی و مهندسی نشان داد با توجه به این‌که ماهیت آزمون‌های دبیرستانی ملاک مرجع می‌باشند اما هنگامی که نمرات سوابق تحصیلی، تنها برای رتبه‌بندی افراد به‌کار می‌رود از پارامترهای مناسبی برخوردار است. درصد بالایی از تغییرات نمرات افراد در دو آزمون عمومی (۶۸٪) و اختصاصی (۷۳٪) مربوط به تفاوت‌های سیستماتیک توانایی افراد p آزمون شونده است تا خطاهای مربوط به مؤلفه‌های هر آزمون t و خطاهای باقیمانده و یا عملیاتی نشده pt,e. برتری نمرات سوابق تحصیلی نسبت به آزمون سراسری این است که برای گروه ریاضی و فنی ۵ درس عمومی و ۵ درس اختصاصی وجود دارد. اگرچه برای یک گروه تخصصی مثل فنی و مهندسی تعداد برابر دروس عمومی و اختصاصی ۵ به ۵ در سوابق تحصیلی همچنان ایراد اساسی دارد اما در مقایسه با آزمون سراسری که نسبت آن ۴ به ۳ به نفع دروس عمومی است از برتری برابری دروس برخوردار است. درنتیجه واریانس نمره‌ی جهانیِ واقعی‌تری به‌دست می‌دهند (به جدول (۲) نگاه کنید).

به‌هرحال انتظار این است، وقتی یک رشته عنوان تخصصی یدک می‌کشد دست‌کم تعداد دروس اختصاصی آن، بیش از دروس عمومی همان رشته باشد. این مشکل زمانی بیشتر می‌شود که بدانیم در آزمون کنکور افراد این گروه عملاً ۴ نمره از دروس عمومی دریافت می‌کنند و ۳ نمره از آزمون‌های اختصاصی که به‌صورت یک نمره گزارش می‌شود و سپس با جمع و ضرب و تبدیل

خطی نمرات و یا به‌عبارتی با دادن ضریب اسمی به آزمون‌ها مشکل نابرابری وزن دروس اختصاصی با دروس عمومی را حل می‌کنند.

در پاسخ به سؤال دوم این پژوهش مبنی بر «۲- اعتبار نمرات مربوط به آزمون سراسری با توجه به میزان خطای آزمون‌های (عمومی و اختصاصی) مختلف r، چقدر است؟ G- study»

این وضعیت برای نمرات کنکور هم کم‌وبیش درست است یعنی هنگامی‌که نمرات آزمون سراسری برای رتبه‌بندی افراد به‌کار می‌رود از پارامترهای مناسبی برخوردار است. به‌عبارتی درصد بالایی از تغییرات نمرات افراد در دو آزمون عمومی (۵۴٪) و اختصاصی (۷۳٪) مربوط به تفاوت‌های سیستماتیک افراد p آزمون‌شونده است تا خطاهای مربوط به مؤلفه‌های هر آزمون t و خطاهای باقیمانده و یا عملیاتی نشده pt,e (به جدول (۳) نگاه کنید). اما ایراد آزمون سراسری این است که برای دروس عمومی ۴ سطح (۱-عربی، ۲-دینی و معارف، ۳- فارسی و ۴- زبان) و برای دروس اختصاصی ۳ سطح تعریف کرده‌اند که پس از ضریب اسمی دادن به دروس اختصاصی وزن آن‌ها را در مجموعه‌ی آزمون نسبت به آزمون‌های عمومی افزایش می‌دهند.

این در حالی است که در مجموعه‌ی دروس اختصاصی این گروه آزمایشی، درس ریاضی از سه مؤلفه‌ی ۱-جبر و احتمال، ۲-هندسه و ۳-حسابان تشکیل شده که نمرات این سه درس جمع می‌شوند و یک نمره از جمع هر سه درس می‌سازند. به‌عنوان مثال دو فرد با دو شیوه‌ی متفاوت که یکی از آن‌ها به ۵ سؤال هندسه درست جواب می‌دهد و نفر بعدی که به ۵ سؤال حسابان درست جواب می‌دهد می‌توانند نمره‌ای مشابه دریافت کنند که از لحاظ آماری به همان ترتیبی که توضیح داده شد و اصول سنجش و اندازه‌گیری و ارزش‌یابی آموزشی درست نیست. این ایراد در بعضی از مواد آزمون‌های دکترا نیز وجود دارد. بر اساس منطق رگرسیون، با افزایش تعداد آزمون x می‌توان پیش‌بینی دقیق‌تری از پارامتر توانایی Y جامعه آزمون‌شوندگان به‌دست آورد و آن‌ها را بهتر تفکیک کرد. بنابراین برای گروه ریاضی و فنی بهتر است این سه درس ۱-جبر و احتمال، ۲-هندسه و ۳-حسابان تفکیک و تبدیل به سه نمره‌ی جداگانه شوند. تعداد سؤالات هرکدام از این سه درس مشخص و در طول دوران ثابت و بدون تغییر بماند. برای رشته‌های مختلف دانشگاهی سهم هرکدام از این دروس و یا به‌عبارتی وزن این سه درس ریاضی و دروس دیگر تخصصی تعیین شود؛ اگر سه درس باهم برابر باشند-مثل آزمون ریاضی همین تحلیل- و در مجموعه‌ی آزمون هر سه درس باهم ضریب ۴ دارند، با دادن ضریب اسمی ۱/۳۳ به هر درس وزن آن‌ها را مشخص و تفکیک کنند. سپس از معادله‌ی رگرسیون نمرات را به‌دست آورند نه سرجمع ساده‌ی نمرات دروس. پس از تعیین همه‌ی جزئیات، هرساله از یک الگوی معین و مشخصِ کارشناسی شده استفاده شود.

در پاسخ‌گویی به سؤال سوم این تحقیق «۳-آیا با ترکیب کردن نمره‌ی کنکور به سوابق تحصیلی می‌توان نمره‌ی معتبری به‌دست آورد؟ G- study»

برای داده‌های ترکیبی حاصل از سوابق تحصیلی و آزمون سراسری که به‌صورت طرح تعمیم‌پذیری چندمتغیره است واریانس فرد، سطح (آزمون عمومی و اختصاصی) و لایه‌های هر سطح (آزمون‌های منفرد) به همراه تعامل دوطرفه آن‌ها و خطاهای غیرسیستماتیک در جدول (۶) برآورد شده است. درصد واریانس هر مؤلفه، نشان می‌دهد که واریانس نمرات افراد پیش از آن‌که برآمده از نمره‌ی جهانی (یا نمره واقعی) آن‌ها باشد نتیجه‌ی تغییرات ناشی از لایه‌های t آزمون‌های عمومی و اختصاصی r است یعنی t: r می‌باشد. به‌عبارت دیگر هر آزمون منفرد t که در دل آزمون‌های عمومی یا اختصاصی r هم‌آشیان شده، مقدار واریانسی تولید می‌کنند که از واریانس تفاوت واقعی توانایی آزمون‌شوندگان p بیشتر است. در تحلیل واریانس این مشکل هنگامی پیش می‌آید که تفاوت‌های درون‌گروهی SSw بزرگ‌تر از تفاوت‌های بین‌گروهی SSb باشد.

مشکل نابرابری تعداد دروس عمومی و اختصاصی هم در سطوح و هم در لایه‌های هر درس در این‌جا نیز وجود دارد؛ نابرابری سطوح r به‌ویژه هنگامی بیشتر نمود پیدا می‌کند که بدانیم برای یک رشته‌ی تخصصی مثل گروه ریاضی به علت دارا بودن ۴ بند سطوح آزمون عمومی در مقابل ۳ بند سطوح آزمون اختصاصی وزنی برابر با w= ۰/۵۳ برای دروس عمومی ایجاد می‌کند. این به معنای آن است که برای فردی که در رشته‌ی ریاضی است عملاً وزن مؤثر آزمون‌های اختصاصی برای او ۳٪ کمتر از وزن آزمون‌های عمومی است. اگرچه دروس عمومی به علت تعداد بیشتر آزمون‌های آن، وزن w بیشتری نیز دارند اما سهم درصدی ٪ آن‌ها از کل واریانس برابر با ۴۴/۳۷٪ است؛ در مقابل سهم درصدی آزمون‌های اختصاصی که برابر با ۵۵/۶۳٪ است. به‌عبارت بهتر برای گروه تخصصی ریاضی و فنی تعداد آزمون عمومی، بیشتر (هزینه و زمان بیشتر برای نظام آموزشی و آزمون‌شوندگان) و خاصیت کمتری دارد.

همچنین این نابرابری تعداد سطوح در آزمون سراسری مشکل ترکیب دو نوع نمره یعنی نمرات سوابق تحصیلی و نمرات آزمون سراسری برای ساخت نمره‌ی کل جهانی را چند برابر می‌کند. البته همان‌طور که گفته شد در سازمان سنجش برای درس‌های تخصصی وزن‌های اسمی (یا ضرایب) متفاوتی برای رشته‌های مختلف در نظر گرفته می‌شود که این امر تا حدودی کمک می‌کند اثر وزن مؤثر دروس نامربوط با اهداف رشته کاهش پیدا کند (ذوالفقارنسب، خدایی و یادگارزاده، ۱۳۹۲). بااین‌حال در یک آزمون مربوط به یک رشته‌ی تخصصی چنین وضعیتی اتلاف منابع ملی به‌ویژه هدر دادن وقت و زمان آزمون‌شوندگان است. بهترین حالت زمانی است که از سهم چنین دروسی کاسته و به نفع دروس اختصاصی بعضی لایه‌های دروس عمومی برای امتحان

حذف کامل شوند و یا در طراحی محتوای آموزشی بعضی سری کتاب‌ها تجدیدنظر شود و «دست‌کم برای کاهش هزینه‌های ملی» تبدیل به یک کتاب با حجم و محتوای کمتر شوند مثل مجموعه دروس عربی با حجم مطالب و ساعات آموزشی کمتر. به‌هرحال نویسندگان معتقد به کاهش حجم محتوای کتب درسی عمومی در راستای اثربخش کردن آن‌ها برای اهداف توسعه‌ی ملی هستند و دروس ذکر شده به‌ویژه درس فارسی تنها به‌عنوان مثال بیان شده‌اند. وقتی برای دروس تخصصی یک رشته ضریب اسمی تعیین می‌شود درواقع هزینه‌های مادی و معنوی هدررفته‌ی دروس نامربوط را با ضرب و تقسیم ساده جبران می‌کنیم. این درحالی است که در بسیاری از کشورهای درحال‌توسعه که اصلاحات موفقی در نظام‌های آموزشی خود پایه‌ریزی کرده‌اند هم‌پوشانی و تکرار بین محتوای دروسِ بااهمیت کمتر آن‌ها کاسته شده است (ذوالفقارنسب، شادمهر و نقی‌زاده، ۱۳۹۲).

وقتی آزمون‌های ملاک مرجع با آزمون‌های هنجار مرجع برای ساخت یک نمره در یک سطح کلی‌تر (عمومی- اختصاصی) ترکیب می‌شوند، عمده واریانسی که به‌دست می‌دهند ناشی از تعامل این نوع آزمون‌ها با سطوح کلی‌تر (عمومی- اختصاصی) t: r است تا واریانس واقعی مربوط به تفاوت افراد p. عمدتاً سوابق تحصیلی از نمره‌ی آزمون‌های ملاک مرجعی تشکیل شده که توزیع نمره‌های آن‌ها چولگی چپ دارد. برعکس توزیع نمره‌ی خام آزمون‌های هنجار مرجع کنکور به علت سختی سؤالات آن، برای پاسخ‌گویی درست دارای چولگی شدید راست است. ترکیب این دو نوع نمره باعث ایجاد یک توزیع دو نمایی می‌شود که فرض خطی بودن رابطه بین توانایی x و نمره‌ی پیش‌بینی شده‌ی Y را بر اساس مدل رگرسیون مخدوش می‌سازد.

از طرف دیگر ماهیت و هدف غایی هر دو نوع آزمون با یکدیگر تفاوت دارد. امتحانات دوره‌های تحصیلی عمدتاً برآمده از محتوای آموزشی است و ارتباط تنگاتنگی دارد با آن‌چه به‌طور مستقیم در کلاس درس آموزش داده شده است. برعکس آزمون سراسری از انعطاف و تنوع بیشتری برخوردار است و برای این‌که ماهیت بردوباخت در یک ماراتن (کنکور) را دارد می‌تواند از دروس آموزش داده شده در دوره‌ی آموزشی فاصله بگیرد. این‌ها عواملی هستند که در ظاهر جفت‌وجور کردن نمرات این دو نوع آزمون را با مشکل بی‌اعتباری مواجه ساخته است. بااین‌حال تبدیل نمرات این دو نوع آزمون به نمرات z استاندارد می‌تواند بخشی از مشکلات را به‌خوبی حل کند. اگرچه برای کاهش چولگی قاعده این است که ابتدا نمرات تبدیل به توزیع صدکی شوند چون اگر مستقیماً تبدیل به توزیع z شوند چولگی نمرات از بین نمی‌رود؛ سپس برای قابل مقایسه کردن نمرات آزمون‌های مختلف تبدیل به توزیع z استاندارد شوند. این کار کمک می‌کند نمرات دارای میانگین و واریانس یک‌سانی شوند و آن‌ها را به بهترین صورت ترکیب کرد (مگنسون،

۱۹۶۶). به‌علاوه تنها مفروضه‌ی تئوری تعمیم‌پذیری که در آن فرض می‌شود آزمون‌ها یا سؤالات از جهانی مشابه از آزمون‌ها یا سؤالات موازی بیرون آمده‌اند به‌طور ضمنی پذیرفته شود (براون، ۲۰۰۵).

بر اساس جدول (۶) با این‌که مقدار ضریب تعمیم‌پذیری و شاخص اعتمادپذیری به‌دست‌آمده برای ترکیب این دو نوع نمره یعنی سوابق تحصیلی با آزمون سراسری بالا است اما همچنان از درون یک نابرابری و ترکیب واریانس‌های ناخواسته حاصل از تفاوت نوع آزمون‌ها یا به‌عبارتی لایه‌ها به‌دست‌آمده که اگرچه لایه‌های دروس از نظر محتوا از جهانی مشابه نمونه‌گیری شده‌اند اما موازی نیستند: یعنی ترکیب نمرات آزمون‌های هنجار مرجع با آزمون‌های ملاک مرجع که در سطرهای بالا توضیح داده شده‌اند (در جدول (۶) به درصد واریانس واقعی p که ۲٪ و ۶٪ است نگاه کنید!). به‌عبارت‌دیگر ممکن است ضرایب تعمیم‌پذیری از واریانس‌های ناخواسته‌ای مثل تغییرات ناشی از هر سطح و لایه‌های آن (فرم‌های آزمون‌ها (t به‌دست‌آمده باشند نه واریانس نمره‌ی جهانی p. بنابراین در تحلیل‌هایی که بر اساس تئوری تعمیم‌پذیری صورت می‌پذیرد علاوه بر این‌که ضرایب $E\hat{p}^2_{\delta}$ باید بزرگ باشد واریانس p نیز باید زیاد باشد تا بتوان ثابت کرد که بالا بودن ضرایب تعمیم‌پذیری نمره‌ی جهانی افراد عمدتاً ناشی از تغییرات سیستماتیک نمره‌ی واقعی p آن‌ها است نه عوامل دیگر. درنهایت در یک طرح تعمیم‌پذیری، تنها دست‌یابی به یک ضریب تعمیم‌پذیری و یا شاخص اعتمادپذیری بالا از اهمیت چندانی برخوردار نیست. درصد واریانس نمره‌ی جهانی p که نشان‌دهنده‌ی میزان تغییرات نمره‌ی واقعی افراد از یکدیگر است برای اعتبار تعمیم یک نمره‌ی جهانی از اهمیت بیشتری برخوردار است (مگنسون، ۱۹۶۶).

«۴- افزایش یا کاهش تعداد آزمون‌های مختلف چه تأثیری بر اعتبار اندازه‌گیری می‌گذارد؟ «D-study

بر اساس جدول (۸) برابر کردن سطوح آزمون عمومی r_1 و اختصاصی r_2 یعنی ۳ به ۳ همچنین تفکیک کردن نمرات حاصل از دروس ریاضی t به ۴ نمره جدا از هم باعث کاهش واریانس خطای سطوح و لایه‌ها خواهد شد. بدین ترتیب که واریانس نمره جهانی p برای آزمون‌های عمومی به ۱۴٪ و برای دروس اختصاصی به ۳۹٪ افزایش می‌یابد. اگرچه این مقدار نسبت به خطای حاصل از لایه‌های آزمون t: r درصد قابل‌توجهی نیستند اما با حذف یک سطح آزمون عمومی و تفکیک ۴ لایه نمره برای درس ریاضی درصد واریانس نمره‌ی جهانی دست‌کم ۷ برابر افزایش پیدا می‌کند: به ترتیب از ۲٪ و ۶٪ در طرح نابرابر سطوح (جدول ۶) به ۱۴٪ و ۳۹٪ در طرح برابر سطوح با افزایش لایه‌های ریاضی به ۴ درس مجزا افزایش می‌یابد (جدول ۸).

درنهایت استفاده از سوابق تحصیلی به‌تنهایی برای انتخاب افراد کافی نیست و نیاز به نمرات تکمیلی مثل آزمون ورودی سراسری استانداردی است که بتوان هم‌جهت با نمرات سوابق تحصیلی افراد را برای سطوح بالاتر انتخاب کرد. اما شیوه‌ی دوگانه‌ی آزمون‌گیری دبیرستان و کنکور ورودی دانشگاه ترکیب نمرات آن‌ها را کم‌اعتبار می‌سازد. هدف غایی هر دو آزمون، دادن امتیاز به افراد توانمندتر برای ادامه تحصیل به سطوح بالاتر و درنتیجه تهیه‌ی منابع انسانی و نیروی کار چرخه‌ی اقتصاد و پیشبرد توسعه‌ی ملی است اما با دو رویکرد متفاوت (هنجار مرجع در مقابل ملاک مرجع). شایسته است این دو رویکرد دست‌کم در بعضی نقاط مشترک به یکدیگر بپیوندند. به دنبال این دو رویکرد متفاوت مؤسسات آموزشی آمادگی کنکور متنوعی ایجاد شده‌اند که شیوه‌های گوناگونی را برای یادگیری (تست‌زنی) تهیه و تدارک دیده‌اند. پیش از هر تغییری هم سازمان سنجش به‌عنوان گلوگاه نظام آموزشی کشور و هم آموزش‌وپرورش و هم آموزش‌عالی باید رویکرد یکپارچه و منسجمی برای هم‌خوانی و هماهنگی بین این شیوه‌های ارتقای افراد در نظام آموزشی کشور سامان دهند. این سه نیازمند تعامل سازنده و همکاری با سازمان‌های مختلف نیز هستند تا بتوانند برای توسعه پایدار مبتنی بر عدالت آموزشی سیاست‌گذاری‌های کلان را جهت‌دهی کنند و هم‌زمان منابع انسانی و نیروی کار کشور را تهیه کنند (ذوالفقارنسب و یادگارزاده، ۱۳۹۱).

یادداشت‌ها

1. قانون فوق مشتمل بر یازده ماده و هفت تبصره در جلسه‌ی علنی روز یک‌شنبه مورخ ۱۳۹۲/۶/۱۰ مجلس شورای اسلامی تصویب شد و در تاریخ ۱۳۹۲/۶/۲۰ به تأیید شورای نگهبان رسید. این قانون از سال تحصیلی (۱۳۹۴ ـ ۱۳۹۳) در صورت اجرائی شدن تبصره‌ی (۵) ماده‌ی (۵) لازم‌الاجرا می‌باشد.

2. criterion-referenced tests
3. norm referenced tests
4. Bond
5. American Educational Research Association
6. composite score
7. generalizability theory
8. liberalization of reliability theory
9. generalizability coefficient
10. index of dependability
11. accuracy of generalization
12. universe score
13. facet
14. conditions
15. fixed effects
16. random effects
17. grand mean
18. universe of items
19. unsystematic or unmeasured error
20. domain referenced
21. rank order
22. generalizability
23. dependability
24. decision study
25. universe score
26. total score metric
27. mean score metric

منابع

الف. فارسی

سرمد، زهره، بازرگان، عباس و حجازی الهه. (۱۳۸۳). *روش‌های تحقیق در علوم رفتاری*، چاپ دهم، مؤسسه انتشارات آگاه.

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

ذوالفقارنسب، سلیمان و یادگارزاده، غلامرضا. (۱۳۹۱). *چهارچوب سامانه نوآوری و کارکرد آن در رشد و توسعه ملی: رویکردی بر پایه نشان‌گرهای سازمان ملل متحد*، رهیافت ۵۴، ص، ۴۱-۵۰.

ذوالفقارنسب، سلیمان، خدایی، ابراهیم و یادگارزاده، غلامرضا. (۱۳۹۲). وزن دهی بهینه به سؤال‌ها و خرده آزمون‌های ورودی برای ساخت نمره کل ترکیبی، *فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی*، سال سوم، شماره ۴، ص. ۱۰۴-۷۹.

ذوالفقارنسب، سلیمان، شادمهر، عبدالکریم و نقی‌زاده، سیما. (۱۳۹۲). پیمایش عملکرد استانی گروه آزمایشی ریاضی-فنی آزمون سراسری در چهارچوب توسعه ملی، *فصلنامه سیاست علم و فناوری*، سال ششم، شماره ۲، ص.۱۰۹-۹۳.

مگنسون، داوید. (۱۹۶۶). *مبانی نظری آزمون‌های روانی*. ترجمه محمد نقی براهنی، ناشر دانشگاه تهران ۱۳۵۱.

ب. انگلیسی

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

Bond, L. A. (1996). Norm-and Criterion-Referenced Testing. ERIC/AE Digest.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, 11(4), 27-34.

Brennan, R. L. (2003). Coefficients and indices in generalizability theory. *Center for Advanced Studies in Measurement and Assessment, CASMA Research Report*, 1, 1-44.

Brennan, R. L. (2001). Generalizability theory: Statistics for social science and public policy. *New York: Springer-Verlag. Retrieved March, 30*, 2013.

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24(1), 1-21.

Brown, J. D. (2005). Statistics corner, questions and answers about language testing statistics: Generalizability and decision studies. *Shiken: JALT Testing & Evaluation SIG Newsletter*, 9(1), 12-16.

Cohen, L. G. & Spenciner, L. J. (1998). *Assessment of Children and Youth*. Longman, A Division of Addison Wesley Longman, Inc., 1900 East Lake Avenue, Glenview, IL 60025.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). The Dependability of Behavioral. *Measurements. New York: John Wiley*.

He, Q. (2009). Estimating the reliability of composite scores.

Mushquash, C. & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analyses. *Behavior research methods*, 38(3), 542-547.

Nußbaum, A. (1984). Multivariate generalizability theory in educational measurement: An empirical study. *Applied Psychological Measurement*, 8(2), 219-230.

Ødegård, A., Hagtvet, K. A. & Bjørkly, S. (2008). Applying aspects of generalizability theory in preliminary validation of the Multifacet Interprofessional Collaboration Model (PINCOM). *International Journal of Integrated Care*, *8*(4).

Shavelson, R. J. & Webb, N. M. (1981). Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology*, *34*(2), 133-166.

Shavelson, R. J. & Webb, N. M. (2006). Generalizability theory. *Handbook of complementary methods in education research*, 309-322.

Shavelson, R. J., Webb, N. M. & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, 44(6), 922.

Shavelson, R. J. & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.

VanLeeuwen, D. M., Dormody, T. J. & Seevers, B. S. (1999). Assessing the reliability of student evaluations of teaching (SETS) with generalizability Theory. *Journal of Agricultural Education*, 40, 1-9.

Webb, N. M., Shavelson, R. J. & Haertel, E. H. (2006). 4 Reliability Coefficients and Generalizability Theory. *Handbook of statistics*, 26, 81-124.

Webb, N. M. & Shavelson, R. J. (2005). Generalizability theory: overview. *Wiley StatsRef: Statistics Reference Online*.

Studies in Learning & Instruction

Vol. 8, No. 2, 2017, Ser. 71/2

Investigating the validity of grade point average scores , national university entrance exam scores and composite scores using Generalizability Theory

Soleyman Zolfagharnasab* **Gholamreza Yadegarzadeh**** **Ehsan Jamali***** **Ebrahim Khodayei******

National Organization of Educational Testing

Introduction

According to the ratified law to accept students in future based on increasing the share of their Grade Point Average (GPA) scores in combination with National Entrance Exam (NEE) scores and (omitting the NEE's role gradually) specifying quality and quantity of the combination of these two scores for the good selection of university applicants has been one of the most challenging problems for the decision makers at National Organization of Educational Testing (NOET).

Objective

This research has been done in order to assess the dependability of GPA scores, NEE scores, and their combination based on Generalizability Theory.

Method

The research method was descriptive based on which GPA and NEE scores of 5608 students in mathematics and engineering were analyzed.

Results

Results have shown that each of dual testing methods (GPA scores and NEE scores) used for ranking and selecting the cognizant applicants to the next educational level has the optimum generalizability coefficient; The true variance of each set of these scores can rank persons for the next level of education, but the combination of the two was not efficient .

Conclusion

The NEE is a norm-referenced test with positive skewness, and GPA scores coming from different test are criterion-referenced tests with negative skewness .Combination of these scores has made a bi-modal distribution which produces zero or negative variance and makes the generalizability coefficient invalid. Converting scores to percentile ranks for elimination skewness and then converting ranks to standard-z score distribution in order to combine the distributions of these scores results in increasing the true variance and improving both generalizability coefficients and Index of dependability .Moreover ,to composite score for this group the levels of general exam in NEE should be reduced to 3

*(Corresponding author) Research expert, Research and evaluation center. National Organization of Educational Testing, Iran .Email :salarnik2001@yahoo.com.

**Assistant professor, National Organization of Educational Testing. Iran.

***Assistant professor, National Organization of Educational Testing. Iran.

****Minister's deputy and the hed of National Organization of Educational Testing. Iran

produce a total scores and simultaneously the unique score of mathematic sub-tests in specialized test-split to 4 layers of scores.

Keywords: Generalizability Theory ,Measurement errors ,Universe score ,True variance.