فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی
سال اول، شمارهٔ ۱، بهار ۱۳۹۱، صص ٦٨- ٤٩

Journal of Measurement & Educational Evaluation
Vol. 1, No. 1, Spring 2012

مقایسه روش‌های تخصیص وزن‌های بهینه به مؤلفه‌های آزمون‌های مرکب

Comparing the Methods of Assigning Optimum Weights to the Components of Composite Tests

امیدعلی آقابابایی *
دکتر ابراهیم خدایی **
دکتر سیما نقی‌زاده ***

Omidali Aghababaei
Ebrahim Khodaei (Ph.D)
Sima Naghizadeh (Ph.D)

تاریخ دریافت مقاله: ۹۰/۱۰/۲۵

تاریخ پذیرش مقاله: ۹۰/۱۲/۰۵

چکیده: یک مسئله مهم در تعیین نمره یک آزمون مرکب، نحوه ترکیب نمره بخش‌های مختلف آزمون یا اصطلاحاً خرده آزمون‌ها به منظور محاسبه نمره کل داوطلبان است. این وزن‌ها باید به گونه‌ای انتخاب شوند که ضمن در نظر گرفتن ویژگی‌های روان‌سنجی هر خرده آزمون و مؤلفه‌های تعیین‌کننده آنها، اختلاف نمره مشاهده شده هر داوطلب با نمره واقعی که بیانگر توانایی واقعی داوطلب است را حداقل نماید. به عبارت دیگر چارچوب تصمیم‌گیری بر مبنای ملاحظات مختلف نظیر روایی محتوا (اعتبار)، دفعات آزمون، پایایی و ... طراحی می‌شود.

برای به دست آوردن حداکثر پایایی در نمره مرکب روش‌های مختلفی در چند دهه اخیر پیشنهاد شده است. این روش‌ها به دو دسته کلی، روش‌های ضمنی و عینی تقسیم می‌شوند. روش‌های ضمنی شامل جمع کردن نمرات خام و مدل‌بندی سئوال- پاسخ و روش‌های عینی شامل وزنی کردن مؤلفه‌ها به وسیله میزان دشواری سئوالات، تخصیص وزن‌ها به نمرات مؤلفه‌ها بر مبنای ماکسیمم اندازه پایایی، و وزنی کردن مؤلفه‌ها بر اساس ماکسیمم کردن روایی نمرات مرکب است. در این مقاله، رویکرد به دست آوردن حداکثر پایایی را در دو نظریه کلاسیک آزمون‌سازی و نظریه سئوال- پاسخ معرفی و ضمن بیان نقاط ضعف و قوت هر روش، برآورد پایایی نمره مرکب و خطای استاندارد اندازه‌گیری در این روش‌ها برای داده‌های شبیه‌سازی شده بررسی خواهیم کرد.

واژگان کلیدی: نظریه کلاسیک آزمون‌سازی، نظریه سئوال- پاسخ، نمرات مرکب، پایایی

Abstract: An important issue in obtaining the score in a composite test is the status of combining to combine different component scores of the test to compute the total scores of examinees. These weights should be selected in a manner that not only considers the psychometric properties of each component and their determining elements, but also minimizes the difference between the observed score and the real score of each examinee which explains his/her real ability. In other words, the framework of decision-making is designed with respect to different considerations such as validity, test times, reliability, etc.

There have been suggestions for obtaining to obtain the maximum reliability of composite scores in the last few decades. These include the implicit approach and the explicit approach. The implicit approach involves adding the raw scores and using IRT model. The explicit approach involves weighting the components by the difficulty of the items, assigning the weights to component scores based on the reliability measures of the components, and weighting the components by maximizing the validity of the composite scores. In this paper, we introduce the approach of obtaining the maximum reliability in Classical Test Theory and Item Response Theory. Besides considering the pros and cons of each method, we investigate the estimates of the reliability and the standard error of measurement of the composite scores for data in a simulation study.

Key words: Classical Test Theory, Item Response Theory, Composite scores, Reliability

* پژوهشگر سازمان سنجش آموزش کشور (مسئول مکاتبات: omidali_221@yahoo.com)
** دانشیار سازمان سنجش آموزش کشور
*** استادیار سازمان سنجش آموزش کشور

مقدمه

یکی از مسائل مهم محققان در بخش آزمون‌سازی، وزن‌دهی به مؤلفه‌ها یا خرده آزمون‌ها در یک آزمون مرکب است. در این آزمون‌ها، به منظور به دست آوردن نمره مرکب حاصل، وزن سئوالات یا بخش‌های مختلف این آزمون، معمولاً بر اساس سطح دشواری، میزان اهمیت سئوال یا عوامل دیگر تعیین می‌شوند. بسیاری از آزمون‌ها در سازمان سنجش آموزش کشور و مؤسسه بین‌المللی ETS[1] از این دسته از آزمون‌ها هستند.

در آزمون‌های سازمان سنجش نظیر آزمون سراسری، آزمون کارشناسی ارشد و ... ضرایب مواد امتحانی برای رشته‌ها و گرایش‌های مختلف با یکدیگر متفاوت است. به عنوان مثال در آزمون سراسری برای رشته علوم ریاضی و فنی، نمره کل از ترکیب نمرات دروس عمومی و تخصصی به دست می‌آید. در اینجا منظور از وزن‌دهی، استفاده از وزن‌دهی افتراقی[2] به جای وزن‌دهی ثابت است. هنگامی که تأثیر سئوالات یا خرده آزمون‌ها را بر نمره کل به شکل متفاوت بخواهیم، چندین روش برای وزن‌دهی به خرده آزمون‌ها وجود دارد. بعضی از این روش‌ها، بر مبنای مفاهیم آماری و بعضی دیگر بر اساس یک ملاک بیرونی است.

ترکیب نمرات به منظور تشکیل یک نمره مرکب ممکن است در سطوح مختلف انجام شود. در بالاترین سطح، وزن‌ها از ترکیب نمرات از آزمودنی‌های مختلف با مقیاس‌های مختلف به دست می‌آیند. به طور تجربی وزن‌های حاصل از معیارهای مشخص، به منظور بهینه شدن نمره مرکب به دست می‌آیند. یکی از این معیارها، پایایی[3] نمره مرکب است. منظور از پایایی نمرات آزمون این است که با تکرار آزمون، نتایج حاصل در شرایط یکسان تغییر نکند. خطای وابسته به نمرات مرکب در صورت پایا بودن وزن مؤلفه‌ها یا خرده آزمون‌ها کاهش می‌یابد. بنابراین در این روش به مؤلفه‌هایی که پایایی بیشتری دارند، وزن‌های بیشتری تعلق می‌گیرد.

مقدار پایایی آزمون تابع تعداد سئوالات آزمون است. می‌توان نشان داد که هنگامی که طول خرده آزمون دو برابر می‌شود، واریانس واقعی آن چهار برابر واریانس آزمون اولیه می‌شود. در حالیکه واریانس خطا به نسبت افزایش طول خرده آزمون افزایش می‌یابد، ضریب پایایی نسبتی از واریانس کل آزمون است که واریانس واقعی است.

1. Educational Testing Service
2. Differential Weighting
3. Reliability

علاوه بر افزایش تعداد سئوالات خرده آزمون، عوامل دیگری در افزایش پایایی مؤثر هستند که از آن جمله، افزایش تعداد مقوله‌های محتوایی خرده آزمون و یا خود آزمون و در نهایت انتخاب سئوالاتی که بیشترین همبستگی را با مجموعه سئوالات دارند. باید توجه داشت هنگامی که آزمون روی یک نمونه آزمون‌دهنده‌ها نامتجانس است، پایایی آزمون افزایش می‌یابد. در حالیکه برای نمونه‌های متجانس و همسان، ضریب پایایی کاهش می‌یابد.

تحقیقات زیادی در دهه گذشته به منظور یافتن وزن‌های بهینه در آزمون‌هایی که دارای خرده آزمون هستند، انجام شده است. راندر[1] (۲۰۰۱) تخصیص وزن‌ها به نمرات خرده آزمون‌ها را به روش‌های ضمنی و عینی تقسیم کرد. در روش ضمنی دو روش به منظور ترکیب نمرات مؤلفه‌ها وجود دارد. روش اول جمع کردن نمرات خام مؤلفه‌ها به منظور به دست آوردن نمره مرکب است (که معادل مؤلفه‌های وزنی به وسیله نمرات ماکسیمم آنها است). روش دوم استفاده از مدل نظریه سئوال- پاسخ[2] (IRT) به منظور تحلیل پاسخ‌ها از همه مؤلفه‌ها به طور همزمان برای ایجاد سئوال و اندازه‌های انفرادی است. راندر (۲۰۰۱) نتایج جمع کردن نمرات مؤلفه خام برای ویژگی‌های روان‌سنجی شامل روایی[3] و پایایی نمره مرکب را بررسی کرد. او نشان داد که جمع کردن نمرات خام، تشخیص اختلاف در اهمیت مؤلفه‌ها و سئوالات در نمره مرکب را با مشکل مواجه می‌کند.

در حالت استفاده از مدل‌بندی نظریه سئوال- پاسخ به منظور ترکیب نمرات مؤلفه یا اندازه‌های توانایی فرد، سئوالات از همه مؤلفه‌ها به طور همزمان برای برآورد ویژگی‌های سئوالات و اندازه‌های توانایی فرد روی مقیاس توانایی مرکب کالیبره می‌شوند. لرد[4] (۱۹۸۰) روش‌های استفاده از نظریه سئوال- پاسخ را به عنوان وزن‌هایی برای نمره‌دهی بهینه پیشنهاد کرد.

روش‌های عینی برای ترکیب نمرات مرکب، عموماً شامل وزن‌های تخصیص یافته به سئوالات مؤلفه‌های انفرادی خواهد بود، به طوری که برای این منظور از سه روش عینی استفاده می‌شود. در روش اول، وزنی کردن مؤلفه‌ها به وسیله میزان دشواری سئوالات است. این روش امتیاز مازاد برای سئوالات مشکل در نظر می‌گیرد. با این وجود، در این روش برای سئوالات مشکل رها شده نیز جریمه بیشتری در نظر خواهد گرفت.

1. Runder
2. Item Response Theory
3. Validity
4. Lord

روش دوم، وزن‌ها را به نمرات مؤلفه‌ها بر مبنای اندازه پایایی مؤلفه‌ها تخصیص می‌دهد. در این حالت، وزن بیشتر به مؤلفه‌هایی با اندازه پایایی بیشتر تخصیص، و وزن کمتر برای مؤلفه‌هایی با اندازه پایایی کمتر خواهد بود. در این روش، خطای وابسته به نمره مرکب از روش ساده جمع کردن نمرات خام کمتر است. به علاوه تولید نمرات مرکب با پایایی حداکثر، به وسیله وزن‌های اولیه مناسب به مؤلفه‌های مختلف امکان‌پذیر است.

روش سوم، وزنی کردن مؤلفه‌ها بر اساس ماکسیمم کردن روایی نمرات مرکب با استفاده از یک معیار خارجی از قبل مشخص است. در این حالت، همچنین رگرسیون چندگانه از معیار مورد نظر روی نمرات مرکب (یعنی ترکیب خطی نمرات مؤلفه‌ها با استفاده از وزن‌ها) به منظور به دست آوردن نمرات مرکب با حداکثر همبستگی بین معیار خارجی و نمرات مرکب، استفاده می‌کند.

همانطور که راندر (۲۰۰۱)، چایلدز[1] و همکاران (۲۰۰٤)، و فلدت[2] و برنان[3] (۱۹۸۹) پیشنهاد کردند، روش ترکیب نمرات مؤلفه‌های انفرادی، باعث ایجاد تنوع در مسائل روش‌شناسی می‌شود، به طوری که شامل بررسی و تفسیر پایایی و روایی نمرات مرکب است. به عنوان مثال گیل[4] و برملی[5] (۲۰۰۸) از یک مطالعه شبیه‌سازی به منظور بررسی تأثیر پایایی بین نشانگر نمرات واحد روی سازگاری[6] نمره دسته‌بندی بندی شده نمرات مجموع برای سطح A استفاده کردند.

در این مقاله، با استفاده از ایده حداکثر پایایی نمرات مرکب به بررسی روش‌های به دست آوردن حداکثر پایایی در دو نظریه آزمون‌سازی کلاسیک و نظریه سئوال–پاسخ می‌پردازیم. بر این اساس، در بخش بعد ضمن معرفی هر نظریه و روش‌های مربوط به هر یک، خطای استاندارد اندازه‌گیری مربوط به هر یک معرفی می‌شود. در بخش سوم با استفاده از داده‌های شبیه‌سازی به محاسبه برآورد و پایایی نمره مرکب داده‌های مربوط می‌پردازیم. در نهایت در بخش آخر نتیجه‌گیری و چند پیشنهاد ارائه خواهد شد.

1. Childs
2. Feldt
3. Brennan
4. Gill
5. Bramley
6. Consistency

۱. حداکثر پایایی نمرات مرکب در نظریه‌های آزمون‌سازی

تأثیر پایایی و دیگر ویژگی‌های روان‌سنجی مؤلفه‌ها روی پایایی نمرات مرکب به طور گسترده برای یک دامنه معنی‌دار از سنجش‌ها مورد مطالعه قرار گرفته‌اند. سئوال اساسی این است که اگر آزمون از مجموعه‌ای متفاوت از سئوالات تشکیل شده و توزیع نمرات هر آزمون دارای ویژگی‌های مختلف باشد، پایایی نمره کل حاصل از نمرات خرده آزمون‌ها چیست. به علاوه از کدام روش وزن‌دهی بهینه در سه سطح گزینه سئوال، خود سئوال و خرده آزمون‌ها برای به دست آوردن نمره دقیق‌تر استفاده کرد. همانطور که قبلاً اشاره شد، دو نظریه کلاسیک و IRT به منظور مدل‌بندی داده‌های روان‌سنجی مورد استفاده قرار می‌گیرد. این روش‌ها توسط کرونباخ[1] و همکاران (۱۹۷۲)، لرد (۱۹۸۰)، کولن[2] و برنان (۲۰۰٤) و وب[3] و همکاران (۲۰۰۷) مورد بررسی قرار گرفتند و به طور گسترده برای به دست آوردن اندازه پایایی نمرات مرکب با استفاده از نمرات وزنی مؤلفه‌ها استفاده می‌شوند. در ادامه به اختصار به معرفی این دو نظریه خواهیم پرداخت.

۱,۲ نظریه کلاسیک آزمون‌سازی[4]

مطابق این نظریه جمع تعداد پاسخ‌های صحیح X به سئوالات یک آزمون را به عنوان شاخص توانایی یا استعداد یک آزمون‌دهنده به کار می‌برند. روش نمره‌گذاری آزمون‌های پیشرفت تحصیلی نیز عمدتاً به این صورت است که به پاسخ صحیح نمره یک و به پاسخ غلط صفر نسبت می‌دهد. به علاوه جمع ساده پاسخ‌های صحیح هر فرد، نمره مشاهده شده X را نشان می‌دهد. فرض اساسی در نظریه کلاسیک آزمون‌سازی یا نظریه نمره واقعی این است که نمره مشاهده شده X هر فرد از مجموع دو بخش شامل نمره واقعی T که بیانگر میزان توانایی داوطلب است و خطای اندازه‌گیری E که سهم سایر عوامل در نمره فرد است، به صورت $X = T \pm E$ تشکیل شده است.

برآورد نمره واقعی T فرد تحت تأثیر چند عامل از جمله مقدار خطای E است که فرض می‌شود دارای توزیع تصادفی نرمال با میانگین صفر در اجراهای مستقل یک

1. Cronbach
2. Kolen
3. Webb
4. Classical Test Theory

آزمون است. در این نظریه، بر اساس ویژگی‌های توزیع نمرات از جمله مقدار خطای E، روش‌هایی تحت عنوان پایایی آزمون، دقت، صحت و اعتبار نمره مشاهده شده X طراحی شده است.

با توجه به رابطه نمره مشاهده شده و نمره واقعی، واریانس نمره مشاهده شده X به صورت زیر خواهد بود:

$$\sigma_{\mathrm{X}}^2 = \sigma_{\mathrm{T}}^2 + \sigma_{\mathrm{E}}^2 \pm 2\sigma_{T,E} \quad (۱)$$

با توجه به اینکه خطای اندازه‌گیری مورد انتظار برای همه نمرات واقعی یکسان است، کواریانس نمره واقعی و خطای اندازه‌گیری ($\sigma_{T,E}$) برابر صفر خواهد بود و بنابراین واریانس نمره مشاهده شده به صورت $\sigma_{\mathrm{X}}^2 = \sigma_{\mathrm{T}}^2 + \sigma_{\mathrm{E}}^2$ است.

اگر واریانس کل نمره مشاهده شده در هر خرده آزمون برابر با واریانس نمره واقعی T آن باشد، در نتیجه، واریانس خطا برابر صفر است و ضریب پایایی در بیشینه ممکن خود یعنی یک قرار می‌گیرد. به این معنی که کل واریانس برآورد شده، واریانس واقعی است. از طرفی، اگر واریانس خطا به حداکثر ممکن برسد و برابر واریانس کل آزمون شود، در این صورت، ضریب پایایی برابر صفر است. چنین وضعیتی وقتی که تمام نمره‌های مشاهده شده افراد از عناصر خطا تشکیل شده باشند، پدید می‌آید و در نتیجه آزمون به طور کلی ناپایا است. بنابراین می‌توان پایایی را به عنوان خارج قسمت واریانس نمره واقعی T به واریانس نمره مشاهده شده X تعریف کرد.

اگر آزمونی بتواند واریانس زیادی تولید کند یا به عبارت دیگر نمرات مشاهده شده (X) افراد گروه نمونه در هر خرده آزمون دارای واریانس زیادی باشد، چنین آزمونی توانسته تغییرات بیشتری را نشان دهد و به دنبال آن توانسته افراد را بهتر تفکیک کند، به این دلیل است که دامنه پراکندگی زیادی را نشان می‌دهد. اما نمره مشاهده شده آزمون‌هایی که واریانس کمی تولید می‌کنند و به عبارت دیگر دامنه مشاهده شده X افراد در این آزمون‌ها محدود به چند نمره است، قدرت تمیز و تفکیک افراد را ندارند. بنابراین به طور خلاصه در نظریه کلاسیک آزمون‌سازی با افزایش واریانس واقعی آزمون منجر به کاهش خطای اندازه‌گیری و در نتیجه افزایش پایایی می‌شود.

تحقیقات مهمی به منظور مطالعه پایایی نمرات مرکب با استفاده از رویکرد کلاسیک آزمون‌سازی انجام شده است. ونگ[1] و استنلی[2] (۱۹۷۰) فرمولی را برای محاسبه پایایی نمرات مرکب که از نمرات مؤلفه‌ها با وزن‌های عینی تشکیل شده‌اند، ارائه نمودند. همچنین فلدت و برنان (۱۹۸۹) روش‌های مختلف و فرمول‌های نظری برای برآورد پایایی نمره مرکب را پیشنهاد کردند. به علاوه فرمول اسپیرمن- براون تعمیم یافته[3] برای نمرات مرکب و ضریب آلفای طبقه‌بندی شده[4] برای آزمون‌هایی که شامل گروه‌هایی با سئوالات همگن، پایایی نمرات اختلاف، پایایی نمرات پیشگویی شده و نمرات عامل هستند، می‌باشند که در ادامه به معرفی آنها خواهیم پرداخت.

۱،۱،۲ فرمول پایایی مرکب عام[5]

فلدت و برنان (۱۹۸۹) یک قضیه آماری پایه درباره نمرات مرکب که توسط ترکیبات خطی مؤلفه‌های وزنی تشکیل شده است و به منظور مطالعه پایایی نمرات مرکب در چارچوب نظریه کلاسیک آزمون‌سازی استفاده می‌شود، فراهم کردند. نمره مرکب L که از n مؤلفه وزنی به صورت $L=\sum_{i=1}^{n} w_i X_i$ تشکیل شده است، به طوری که در آن X_i نمره مؤلفه iام و w_i وزن تخصیص یافته به مؤلفه iام است.

با فرض اینکه خطاها بین مؤلفه‌ها به طور خطی مستقل باشند، پایایی مرکب به صورت زیر به دست می‌آید (فلدت و برنان ۱۹۸۹، وب و همکاران ۲۰۰۷).

$$r=1-\frac{\sigma_{c,e}^2}{\sigma_c^2}=1-\frac{\sum_{i=1}^{n} w_i^2\sigma_{e,x_i}^2}{\sum_{i=1}^{n} w_i^2\sigma_{e,x_i}^2+\sum_{i=1}^{n}\sum_{j(\neq i)=1}^{n} w_i w_j \sigma_{x_i,x_j}}$$

$$=1-\frac{\sum_{i=1}^{n} w_i^2\left(1-r_i\right)\sigma_{x_i}^2}{\sum_{i=1}^{n} w_i^2\sigma_{x_i}^2+\sum_{i=1}^{n}\sum_{j(\neq i)=1}^{n} w_i w_j r_{i,j}\sigma_{x_i}\sigma_{x_j}}$$

1. Wang
2. Stanley
3. Generalized Spearman-Brown
4. Stratified Coefficient Alpha
5. General Composite Reliability

$$= \frac{\sum_{i=1}^{n} w_i^2 r_i \sigma_{x_i}^2 + \sum_{i=1}^{n} \sum_{j(\neq i)=1}^{n} w_i w_j r_{i,j} \sigma_{x_i} \sigma_{x_j}}{\sum_{i=1}^{n} w_i^2 \sigma_{x_i}^2 + \sum_{i=1}^{n} \sum_{j(\neq i)=1}^{n} w_i w_j r_{i,j} \sigma_{x_i} \sigma_{x_j}} \quad (۲)$$

که در آن r_i میزان پایایی مؤلفه iام، $\sigma_{x_i}^2$ واریانس مؤلفه iام، σ_{e,x_i}^2 واریانس خطای مؤلفه iام، σ_{x_i,x_j} کواریانس بین مؤلفه iام و مؤلفه jام، $r_{i,j}$ همبستگی بین مؤلفه iام و مؤلفه jام، σ_c^2 واریانس مرکب و $\sigma_{c,e}^2$ واریانس خطای نمره مرکب است.

فرمول (۲) نشان می‌دهد که پایایی نمره مرکب تابعی از وزن‌های تخصیصی به اندازه پایایی مؤلفه‌های انفرادی، واریانس نمرات مؤلفه‌ها و همبستگی بین نمرات مؤلفه‌ها است. به منظور بهینه کردن پایایی نمره مرکب می‌توان وزن مؤلفه‌های انفرادی را در فرمول (۲) تغییر داد.

۲،۱،۲ فرمول پایایی مرکب ونگ و استنلی

هنگامی که نمرات هر مؤلفه استاندارد می‌شود، فرمول (۲) به فرمول پایایی مرکب ونگ و استنلی تبدیل می‌شود، به طوری که می‌توان نوشت:

$$r = \frac{\sum_{i=1}^{n} w_i^2 r_i + \sum_{i=1}^{n} \sum_{j(\neq i)=1}^{n} w_i w_j r_{i,j}}{\sum_{i=1}^{n} w_i^2 + \sum_{i=1}^{n} \sum_{j(\neq i)=1}^{n} w_i w_j r_{i,j}} \quad (۳)$$

در حالت خاص هنگامی که دو مؤلفه وجود دارد، پایایی نمره مرکب به صورت زیر است:

$$r = \frac{w_1^2 r_1 + w_2^2 r_2 + 2 w_1 w_2 r_{12}}{w_1^2 + w_2^2 + 2 w_1 w_2 r_{12}} \quad (٤)$$

که در این حالت، کمترین مقدار ممکن برای پایایی مرکب برابر پایایی مؤلفه با قابلیت اعتماد کمتر است. اگر دو مؤلفه همبسته باشد، پایایی مرکب از پایایی هر مؤلفه بیشتر است. اگر پایایی دو مؤلفه با هم برابر باشد، در این صورت اگر نسبت وزن‌های دو مؤلفه برابر یک باشد، مقدار ماکسیمم پایایی مرکب برابر $\frac{r_{12} + r_1}{1 + r_{12}}$ خواهد بود.

۳،۱،۲. فرمول اسپیرمن- براون تعمیم یافته

فرمول اسپیرمن- براون تعمیم یافته یک حالت خاص فرمول (۳) است، به طوری که مؤلفه‌ها به صورت واحدهای آزمون موازی با وزن‌های برابر با یک هستند. پایایی نمره مرکب در این حالت به صورت زیر است.

$$r = \frac{nr_1}{1+(n-1)r_1} \quad (۵)$$

که در آن r_1 پایایی یک واحد آزمون است. هنگامی که پایایی نمره مرکب و پایایی واحدهای موازی معلوم هستند، فرمول (۵) به منظور محاسبه طول مورد نیاز آزمون استفاده می‌شود. به علاوه فرمول اسپیرمن- براون تعمیم یافته، معمولاً در مراحل اولیه آزمایش به منظور بررسی اثر طول آزمون و انواع سئوال روی پایایی آزمون نهایی استفاده می‌شود (فلدت و برنان، ۱۹۸۹).

٤،۱،۲. ضریب آلفای طبقه‌بندی شده

همانطور که فلدت و برنان (۱۹۸۹) نشان دادند، حتی برای یک آزمون انفرادی، اندازه‌گیری ساختار موردنظر بر حسب اندازه سئوالات به ندرت همگن هستند. در بسیاری از موارد، سئوالات در یک آزمون به منظور اندازه‌گیری تفاوت ناچیز در بعد دامنه محتوا گروه‌بندی می‌شوند.

ضریب آلفا (آلفای کرونباخ) یک اندازه سازگاری درونی پایایی است که یکی از مهم‌ترین ضرایب مورد استفاده برای اندازه‌گیری پایایی یک آزمون انفرادی اجرا شده، است. سئوالات در یک آزمون بایستی برای ضریب آلفا، تاو- هم ارز[1] باشند تا در نتیجه برآوردگرهای به دست آمده برای پایایی نااریب باشند (وقتی که سئوالات در یک آزمون تاو- هم ارز هستند، تفاوت بین نمرات واقعی برای هر جفت سئوال، ثابت است و سئوالات دارای واریانس نمره واقعی برابر هستند. اگرچه ممکن است واریانس نمره خطای نابرابر داشته باشند). این مسئله معمولاً به ندرت اتفاق می‌افتد، زیرا علاوه بر قدرت تشخیص برابر برای همه مؤلفه‌های آزمون، مستلزم یک بعدی بودن[2] کل آزمون است.

ضریب آلفای طبقه‌بندی شده که حالتی خاص از فرمول (۲) است، به صورت زیر است (فلدت و برنان، ۱۹۸۹):

$$r_{Strat,\alpha} = 1 - \frac{\sum_i \sigma_i^2 (1-r_i)}{\sigma_c^2} \quad (٦)$$

1. Tau-equivalent
2. Unidimetionality

که در آن $r_{Strat,\alpha}$ پایایی نمره مرکب، r_i پایایی طبقه iام، σ_i^2 واریانس طبقه iام و σ_c^2 واریانس نمرات مرکب است.

علی‌رغم اینکه روش‌های ارائه شده قبلی برای آزمون انفرادی شامل گروه‌هایی از سئوالات همگن است، در صورتی که بتوان اندازه‌های پایایی آزمون‌های انفرادی را برآورد و به هر مؤلفه وزن یکسان داده شود، این روش‌ها را می‌توان برای محاسبه پایایی نمره مرکب آزمون‌های انفرادی که به صورت مجزا اجرا می‌شوند، استفاده کرد.

ضریب آلفای طبقه‌بندی شده به طور گسترده به منظور مطالعه پایایی و نمرات آزمون‌هایی که از سئوالات ناهمگن تشکیل شده‌اند، استفاده می‌شود. به عنوان مثال با استفاده از یک مطالعه شبیه‌سازی، ری[1] (۲۰۰۷) نشان داد که هنگامی که مؤلفه‌های اندازه‌گیری کننده عامل‌های مختلف در داخل خرده آزمون‌ها طبقه‌بندی شوند، آلفای طبقه‌بندی شده و ماکسیمم پایایی، برآوردهایی سازگار برای پایایی مرکب فراهم می‌کنند.

خطای استاندارد اندازه‌گیری نمره مرکب[2] در نظریه کلاسیک آزمون‌سازی، برای همه حالت‌های ذکرشده قبلی به صورت زیر محاسبه می‌شود:

$$SEM_{CTT,c} = \sqrt{1 - r_{CTT,c}}\,\sigma_c \quad (۷)$$

که در آن $r_{CTT,c}$ پایایی مرکب محاسبه شده با استفاده از یکی از روش‌های ارائه شده قبلی و σ_c انحراف استاندارد نمره‌های مرکب است.

۲،۲. نظریه سئوال-پاسخ

محدودیت‌های متعددی در استفاده از روش‌های نظریه کلاسیک آزمون‌سازی به منظور محاسبه پایایی مرکب نمرات آزمون وجود دارد. به طوری که در نظریه کلاسیک، ویژگی‌های سئوال و آزمون نظیر دشواری سئوال، قدرت تشخیص و پایایی به نمونه انتخابی از داوطلبان وابسته است. به علاوه در این روش فرض می‌شود که واریانس خطاهای اندازه‌گیری برای همه داوطلبان یکسان است که همواره این مسئله برقرار نیست (لرد، ۱۹۸۰؛ باند[3] و فاکس[4]، ۲۰۰۷).

1. Ray
2. The Standard Error of Measurement
3. Bond
4. Fox

نظریه سئوال- پاسخ گاهی اوقات به نظریه صفت مکنون[1] (پنهان) نیز مشهور است، زیرا در این نظریه، عملکرد آزمودنی در هر یک از سئوال‌های آزمون به صفت مکنون نسبت داده می‌شود. هدف اصلی از اجرای آزمون در نظریه سئوال- پاسخ این است که مشخص شود هر آزمودنی چه مقدار از صفت مورد سنجش آزمون را دارا است. نظر به اینکه بیشتر پژوهش‌هایی که در زمینه‌های روان‌سنجی و تربیتی انجام می‌شوند با متغیرهایی نظیر توانایی خواندن، حساب کردن، ریاضیات و ... ارتباط دارند، در نظریه سئوال- پاسخ از پارامتر توانایی که با نماد θ (تتا) نشان داده می‌شود، استفاده می‌شود.

همبلتون[2] و همکاران (۱۹۹۱) معتقدند که نظریه سئوال- پاسخ دارای دو ویژگی اساسی است. یکی اینکه عملکرد آزمودنی در آزمون به وسیله مجموعه عواملی که صفت مکنون نامیده می‌شود، قابل پیش‌بینی است. دوم اینکه رابطه بین عملکرد آزمودنی در سئوال و صفت مکنون به وسیله تابع سئوال که منحنی ویژگی سئوال[3] (ICC) نامیده می‌شود، توصیف می‌شود. آگاهی از جایگاه آزمودنی روی مقیاس توانایی به پژوهشگر امکان می‌دهد که احتمال پاسخ‌دهی درست به سئوال را برای او پیش‌بینی نماید، و همچنین اگر پارامتر دشواری سئوال معین باشد از روی آن می‌توان جایگاه آزمودنی را روی مقیاس توانایی تعیین کرد. بنابراین برخلاف نظریه کلاسیک آزمون که در آن پایایی و خطای استاندارد اندازه‌گیری برای تمام نمرات آزمون ثابت در نظر گرفته می‌شد، در نظریه سئوال- پاسخ شاخص دقت اندازه‌گیری برای تمام نمرات آزمون جداگانه برآورد می‌شود. همچنین در این روش، ویژگی‌های سئوال (مانند دشواری سئوال) در آزمون کنترل می‌شوند و مدل‌های نظریه سئوال- پاسخ تفاوت‌های موجود در ویژگی‌های سئوال را کنترل می‌کنند. به علاوه در نظریه سئوال- پاسخ با استفاده از مرجع‌سازی آزمون می‌توان سئوال‌های جدیدی را در آزمون جای داد. پس از آنکه با استفاده از یک مدل نظریه سئوال- پاسخ پارامترهای سئوال برآورد شدند، محقق می‌تواند نمرات قابل مقایسه را بر مبنای یک سازه معین برای آزمودنی‌هایی از جامعه که نتوانسته‌اند به سئوال‌های یکسان پاسخ دهند بدون اقدام به معادل‌سازی محاسبه کند.

1. Latent
2. Hambleton
3. Item Characteristic Curve

مدل‌های مختلف IRT شامل مدل‌های یک پارامتری (مدل راش[1])، دو پارامتری و سه پارامتری برای تحلیل سئوالات دو ارزشی[2]، مدل اعتبار جزئی[3] (PCM) و مدل مقیاس درجه‌بندی[4] (RSM) برای تحلیل سئوالات چند ارزشی[5] است (رایت[6] و مسترز[7]، ۱۹۸۲).

مدل سه پارامتری برای سئوالات دو ارزشی به صورت زیر بیان می‌شود (لرد ۱۹۸۰).

$$P(\theta) = c + (1-c)\frac{\exp\{Da(\theta-b)\}}{1+\exp\{Da(\theta-b)\}} \quad (۸)$$

که در آن θ پارامتر توانایی فرد، $D=1.7$ و $P(\theta)$ احتمال این که یک فرد با توانایی θ به سئوال به درستی پاسخ دهد. a پارامتر تشخیص سئوال، b پارامتر دشواری سئوال و c پارامتر حدس سئوال می‌باشد.

فرمول (۸) نشان می‌دهد که احتمال اینکه یک داوطلب به یک سئوال به درستی پاسخ دهد با افزایش توانایی فرد افزایش و با افزایش دشواری سئوال کاهش می‌یابد. برای $c=0$، این فرمول به مدل راش کاهش می‌یابد. هنگامی که سطح دشواری سئوال و توانایی داوطلب به هم نزدیک است، با استفاده از مدل راش، شانس پاسخ صحیح داوطلب برابر ۰/۵ است.

دو فرض ضروری برای مدل‌های IRT یک متغیره، یک بعدی بودن و استقلال موضعی مدل می‌باشد. یک بعدی بودن مدل مستلزم این است که توانایی یا متغیر پنهان داوطلب توسط آزمون اندازه‌گیری شود. استقلال موضعی نیز مستلزم استقلال آماری پاسخ‌های داوطلبین به هر سئوال آزمون (هنگامی که توانایی آنها بر روی کل آزمون به طور ثابت تأثیر می‌گذارد) می‌باشد. با وجود اینکه این فرض به ندرت در عمل برقرار است، همبلتون و همکاران (۱۹۹۱) نشان دادند که به شرط وجود مقیاس منسجم در سئوالات ساخته شده، فرض یک بعدی بودن همواره الزامی نیست زیرا مدل‌های IRT نسبتاً به این فرض پایدار هستند.

---

1. Rasch model
2. Dichotomous Items
3. Partial Credit Model
4. Rating Scale Model
5. Polytomous Items
6. Wright
7. Masters

استفاده از مدل‌های لجستیک دو پارامتری و سه پارامتری در تفسیر نتایج با مشکلاتی همراه است. به عنوان مثال داوطلبین با نمره خام پایین‌تر ممکن است توانایی برآورده شده بیشتری از افرادی که نمرات خام بالاتر دارند، داشته باشند. با این وجود، بعضی از این محدودیت‌ها، صرفاً تحت شرایطی خاص به وجود می‌آیند. به عنوان مثال در حالی که مدل راش یک ارتباط یکنوا بین اندازه فرد/ سئوال را حفظ می‌کند، مدل اعتبار جزئی صرفاً یک ارتباط یکنوا بین اندازه‌های فرد و نمرات آزمون را حفظ می‌کند. به عبارت دیگر توانایی فرد تابعی از نمرات خام است. این مسئله ممکن است مشکلاتی را در تفسیر برآوردهای مدل اعتبار جزئی به وجود آورد. به طوری که سئوالاتی که درصد نمرات بالاتری دارند (افراد بیشتری به درستی به سئوال- پاسخ می‌دهند). ممکن است از سئوالاتی که درصد نمره پایین‌تری دارند (افراد کمتری به آنها صحیح پاسخ می‌دهند.) مشکل‌تر به نظر آید.

یک مفهوم مهم در مدل‌بندی IRT، تابع اطلاع سئوال است. در حالتی که سئوالات در یک آزمون، دو ارزشی باشند، تابع اطلاع $I_i(\theta)$ برای سئوال iام به صورت زیر تعریف می‌شود (همبلتون و اسوامینان[1]، ۱۹۸۳).

$$I_i(\theta)=\frac{\left(\frac{\partial P(\theta)}{\partial \theta}\right)^2}{P(\theta)\left(1-P(\theta)\right)} \quad (۹)$$

تابع اطلاع آزمون به صورت مجموع اطلاع سئوالات است. به عبارت دیگر

$$I(\theta)=\sum_{i=1}^{n} I_i(\theta) \quad (۱۰)$$

که در آن n تعداد سئوالات آزمون است. برای سئوالات چند ارزشی رایت و مسترز (۱۹۸۲)، و چایلدز و همکاران (۲۰۰٤) مباحث جزئی‌تری را در ارتباط با شکل‌گیری سئوال و توابع اطلاع آزمون و کاربردهای آنها فراهم کردند.

اندازه خطای استاندارد (انحراف معیار) توانایی فرد به صورت معکوس متناسب با اطلاع آزمون است. بنابراین می‌توان نوشت:

$$SEM_{IRT}=\sqrt{\sigma^2_{IRT}}=\sqrt{1/I(\theta)} \quad (۱۱)$$

1. Swaminathan

به دلیل اینکه اطلاع آزمون تابعی از توانایی فرد است، انحراف معیار توانایی فرد نیز تابعی از توانایی داوطلب است. این مسئله متفاوت با رویکرد کلاسیک آزمون‌سازی است که خطای استاندارد اندازه‌گیری عموماً فرض می‌شود که برای همه نقاط نمره یکسان است.

جانسون و جانسون[1] (۲۰۰۹) خطای اندازه‌گیری بر مبنای نظریه کلاسیک آزمون‌سازی و نظریه IRT را مقایسه و بیان کردند که برای مدل IRT ، پارامترهای سئوال از قبل مشخص و ثابت هستند و از این رو خطای اندازه‌گیری اثر عوامل بیرونی نظیر محتوای نمونه‌گیری را منعکس نمی‌کند. با این وجود به دلیل اینکه خطای اندازه‌گیری برای مدل IRT، تابعی از سئوالات در آزمون است، بنابراین می‌توان از آن به عنوان انعکاس اثر نمونه‌گیری از کل سئوالات در برآورد اندازه پایایی داوطلب استفاده کرد.

پایایی مدل راش یا هر مدل IRT دیگر به صورت زیر تعریف می‌شود:

$$R_{IRT} = 1 - \frac{\sigma^2_{IRT,avg}}{\sigma^2_{O,IRT}} \quad (۱۲)$$

که در آن $\sigma^2_{IRT,avg}$ میانگین واریانس خطای اندازه داوطلب و $\sigma^2_{O,IRT}$ واریانس مشاهده شده اندازه داوطلب است.

وقتی که یک آزمون به منظور اندازه‌گیری بیش از یک متغیر پنهان طراحی می‌شود، به طوری که در این حالت آزمون بایستی مقدار مشخصی روایی نظیر محتوای مورد نیاز را داشته باشد، می‌توان از مدل نظریه سئوال- پاسخ چند بعدی[2] (MIRT) استفاده کرد. مدل‌های MIRT به طور خاص برای تشخیص مطالعاتی که به بررسی اینکه چگونه افراد با سئوالات انفرادی با هم اثر متقابل دارند، استفاده می‌شوند (وو[3] و آدامز[4] ۲۰۰۶).

با فرض اینکه هر مؤلفه یک بعد از توانایی داوطلب را محاسبه نماید، توانایی مرکب حاصل از این مؤلفه‌ها برای یک داوطلب θ_c اندازه‌گیری شده به وسیله ترکیب خطی از اندازه‌های توانایی روی مؤلفه‌های انفرادی به صورت $\theta_c = \sum_{i=1}^{n} w_i \theta_i$ به دست آورده می‌شود.

1. Johnson
2. Multidimensional IRT
3. Wu
4. Adams

خطای استاندارد اندازه‌گیری اندازه توانایی مرکب به صورت زیر محاسبه می‌شود:

$$SEM_{IRT,c} = \sqrt{\sigma^2_{IRT,c}} = \sqrt{\sum_{i=1}^{n} w_i^2 \sigma^2_{IRT,i} + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ (j \neq i)}}^{n} w_i w_j r_{ij} \sigma_{IRT,i} \sigma_{IRT,j}} \quad (۱۳)$$

که در آن r_{ij} همبستگی بین اندازه‌های توانایی برای مؤلفه iام و jام، $\sigma^2_{IRT,i}$ واریانس خطای مؤلفه iام است.

در ادامه ضمن مقایسه روش‌های معرفی شده برای به دست آوردن پایایی در نمرات مرکب و خطای استاندارد اندازه‌گیری با استفاده از داده‌های شبیه‌سازی شده، به بیان نتیجه‌گیری خواهیم پرداخت.

۲. شبیه‌سازی

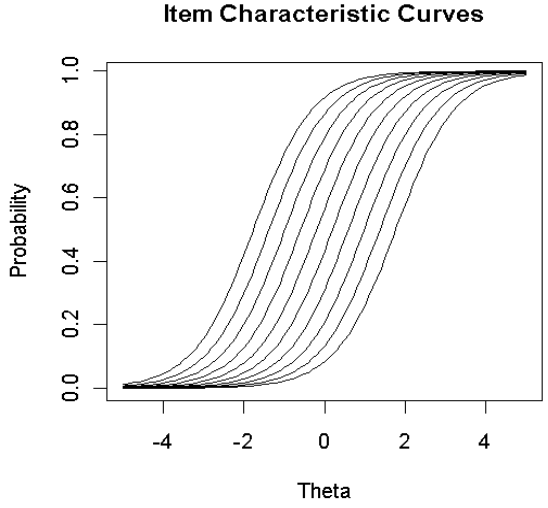
در این بخش، با استفاده از یک مطالعه شبیه‌سازی به برآورد و مقایسه ضرایب پایایی حاصل از روش‌های مختلف معرفی شده در بخش‌های قبل خواهیم پرداخت. برای این منظور با استفاده از نرم افزار WinGen و R برآورد پایایی نمره مرکب و خطای استاندارد اندازه‌گیری در روش نظریه کلاسیک و نظریه سئوال- پاسخ یک بعدی به دست آورده و با یکدیگر مقایسه خواهند شد.

به منظور تولید سئوالات دو ارزشی در مدل سه پارامتری IRT از محیط برنامه‌نویسی در نرم‌افزار R استفاده می‌شود. تعداد داوطلبان در این شبیه‌سازی برای هر دو آزمون ۵۰۰۰ نفر در نظر گرفته می‌شود. همچنین تعداد سئوالات در آزمون اول ٤٠ و برای آزمون دوم ۵۰ در نظر گرفته می‌شود. پارامتر توانایی داوطلبان از توزیع نرمال استاندارد و پارامتر تشخیص از توزیع یکنواخت در بازه ۰/۲ و ۲/۵ تولید می‌شود. به علاوه برای تولید پارامتر توانایی و داده‌های تصادفی از این نرم‌افزار استفاده می‌شود. برای به دست آوردن آماره‌های توصیفی در جدول (۱) و ضرایب پایایی و خطای استاندارد اندازه‌گیری از نرم‌افزارهای WinGen استفاده خواهد می‌شود.

در روش کلاسیک آزمون‌سازی برآورد پایایی نمره کل با استفاده از فرمول پایایی نمره مرکب عام و فرمول ونگ و استنلی برای مؤلفه‌های یک بعدی و چند بعدی داوطلبین استفاده خواهد شد. همچنین خطای استاندارد اندازه‌گیری نیز برای داده‌ها محاسبه خواهند شد.

در نظریه سئوال-پاسخ نیز ابتدا مقادیر توانایی آزمودنی‌ها از توزیع نرمال استاندارد تولید می‌شوند. سپس دو آزمون با یکدیگر ترکیب و با استفاده از مدل‌های معرفی شده تحلیل و اندازه توانایی مرکب و سطح دشواری سئوال، خطای استاندارد اندازه‌گیری به دست آورده می‌شوند. همچنین با کالیبره کردن دو آزمون به طور مجزا، ضریب پایایی توانایی مرکب از روش IRT به دست آورده می‌شود. وزن‌های مختلف به اندازه‌های توانایی مؤلفه‌ها تخصیص و اندازه توانایی مرکب به دست خواهند آمد. در نهایت خطای استاندارد اندازه‌گیری در این روش به دست آورده می‌شود.

نمودار زیر منحنی ویژگی سئوال مربوط به پاسخ‌های تولید شده را به ازای مقادیر مختلف پارامتر θ نمایش می‌دهد. همانطور که ملاحظه می‌شود، با افزایش پارامتر θ یا میزان توانایی داوطلب، احتمال پاسخگویی فرد افزایش می‌یابد. در صورتی که از مدل یک پارامتری IRT استفاده شود، مقدار احتمال پاسخگویی داوطلب به ازای اندازه توانایی صفر برابر ۰/۵ خواهد بود.



جدول (۱) وضعیت توصیفی پاسخ‌های مربوط به هر آزمون را نشان می‌دهد. همانطور که ملاحظه می‌شود، تغییرپذیری نمرات آزمون اول از نمرات دوم بیشتر است. به علاوه ضریب آلفا (کرونباخ) که بیانگر اندازه پایایی داخلی هر آزمون است برای آزمون اول ۰/۹۲۷ و برای آزمون ۰/۸٤٦ در نظر گرفته شده است. همچنین دو آزمون با یکدیگر همبستگی نسبتاً بالایی دارند.

جدول (۱) وضعیت توصیفی پاسخ‌های مربوط به هر آزمون

آماره‌های توصیفی	آزمون اول	آزمون دوم
تعداد سئوالات آزمون	٤٠	٥٠
میانگین نمرات آزمون	٤٢/١٠٤	٣٦/٨٣٩
انحراف معیار آزمون	٥/٤٢٧	٢/٦٠٧٦
ضریب آلفا	٠/٩٢٧	٠/٨٤٦
همبستگی بین دو آزمون	٠/٧٦٨	

جدول (۲) اندازه ضرایب پایایی مرکب و خطای استاندارد اندازه‌گیری را که در بخش‌های قبل به تفصیل معرفی شد، را برای دو روش نظریه کلاسیک آزمون‌سازی و روش نظریه سئوال- پاسخ یک بعدی به ازای ۳ وزن دلخواه برای دو آزمون نمایش می‌دهد. همانطور که ملاحظه می‌شود، در روش نظریه کلاسیک آزمون‌سازی، از فرمول پایایی مرکب عام و فرمول مرکب ونگ و استنلی استفاده شده است. با توجه به همبستگی نسبتاً بالای دو آزمون، اندازه ضرایب پایایی مرکب از ضرایب آلفای مربوط به هر آزمون بالاتر است. از طرفی هنگامی که وزن هر دو آزمون برابر در نظر گرفته می‌شود، ضرایب پایایی برای هر دو روش تقریباً برابر هستند. این مسئله معادل جمع کردن نمرات خام هر دو آزمون است. به علاوه در صورتی که مؤلفه‌های انفرادی دارای ساختار چند بعدی باشند، اندازه پایایی برای هر مؤلفه را می‌توان از ضریب آلفای کرونباخ طبقه‌بندی شده استفاده کرد. همچنین با توجه به اینکه فرمول خطای استاندارد اندازه‌گیری برای هر دو روش یکسان است. بنابراین مقادیر خطا به دست آمده برای هر دو روش برابر خواهد بود.

جدول (۲) مقادیر ضرایب پایایی مرکب و خطای استاندارد اندازه‌گیری برای وزن‌های متفاوت در نظریه کلاسیک و نظریه سئوال- پاسخ

		ضریب پایایی مرکب	خطای استاندارد اندازه‌گیری	وزن آزمون اول	وزن آزمون دوم
روش کلاسیک آزمون‌سازی	فرمول پایایی مرکب عام	٠/٨٨٧	٠/١٨٣	٠/٧٥	٠/٢٥
		٠/٩٢٦	٠/١٩٧	٠/٥٠	٠/٥٠
		٠/٩٥٨	٠/٢٠٤	٠/٢٥	٠/٧٥
	فرمول ونگ و استنلی	٠/٨٩٣	٠/١٨٣	٠/٧٥	٠/٢٥
		٠/٩٢٤	٠/١٩٧	٠/٥٠	٠/٥٠
		٠/٩٤٤	٠/٢٠٤	٠/٢٥	٠/٧٥
روش نظریه سئوال- پاسخ یک بعدی		٠/٨٨٢	٠/١٠٢	٠/٧٥	٠/٢٥
		٠/٩٠٣	٠/١٢٣	٠/٥٠	٠/٥٠
		٠/٩٢٤	٠/١٣٤	٠/٢٥	٠/٧٥

با توجه به اینکه دو آزمون موازی نیستند، بنابراین همانطور که اشاره شد، امکان محاسبه ضریب پایایی مرکب اسپیرمن- براون تعمیم یافته وجود ندارد. با این وجود این ضریب در بیشتر مطالعات به خصوص در مراحل اولیه آزمایش به منظور تعیین اثر طول آزمون روی پایایی آزمون نهایی مفید است. در حالیکه فرمول پایایی مرکب عام و فرمول ونگ و استنلی عموماً به منظور تعیین اثر مؤلفه‌های وزنی روی نمرات مرکب استفاده می‌شوند.

۳. نتیجه‌گیری

در این مقاله به مقایسه رویکرد حداکثر پایایی در وزنی کردن نمرات مرکب در نظریه‌های کلاسیک آزمون‌سازی و سئوال- پاسخ پرداختیم. در نظریه کلاسیک آزمون‌سازی به معرفی فرمول پایایی مرکب عام، فرمول ونگ و استنلی، فرمول اسپیرمن- براون و آلفای طبقه‌بندی به همراه خطای استاندارد اندازه‌گیری پرداختیم. به علاوه با توجه به محدویت‌های حاصل از نظریه کلاسیک نظیر وابستگی آماره‌های سئوال و آزمون به نمونه انتخابی و یکسان بودن واریانس خطای اندازه‌گیری برای همه آزمودنی‌ها، نظریه سئوال- پاسخ بر اساس تعداد پارامترهای موجود در مدل مورد بررسی قرار گرفت. در مطالعه شبیه‌سازی به منظور مقایسه برآورد و پایایی نمره کل داده‌های تولید شده ملاحظه شد که با توجه به تفاوت تغییرپذیری در دو آزمون، پایایی آزمون‌ها نیز به همان صورت متفاوت است. برای روش کلاسیک برآوردها با استفاده از فرمول پایایی مرکب عام و فرمول مرکب ونگ و استنلی به دست آورده که با توجه به همبستگی بالای دو آزمون، برآورد پایایی نمره کل از برآورد هر آزمون بیشتر است. همچنین با توجه به اینکه فرمول خطای استاندارد اندازه‌گیری برای هر دو روش یکسان است. بنابراین مقادیر خطا به دست آمده برای هر دو روش برابر خواهد بود. با توجه به اینکه دو آزمون موازی نیستند، بنابراین همانطور که اشاره شد، امکان محاسبه ضریب پایایی مرکب اسپیرمن- براون تعمیم یافته وجود ندارد.

در روش نظریه کلاسیک آزمون‌سازی، خطای استاندارد اندازه‌گیری برای همه نمرات یکسان در نظر گرفته می‌شود و ضرایب پایایی حاصل اندکی از ضرایب پایایی به دست آمده در روش قبل کمتر است. البته بایستی در نظر داشت که در عمل در نظریه سئوال- پاسخ، برای وزنی کردن دو آزمون از اندازه‌های توانایی استفاده می‌شود در حالیکه در روش قبل، همانطور که اشاره شد، از نمرات خام یا میانگین نمرات استفاده می‌شود.

با توجه به انعطاف‌پذیری بیشتر روش‌های نظریه سئوال- پاسخ نسبت به روش کلاسیک، پیشنهاد می‌شود از این روش در مطالعات و محاسبات در سازمان سنجش مورد استفاده قرار گیرد.

برای مطالعات آینده استفاده از مدل‌های آمیخته را با توجه به یکسان نبودن توزیع نمرات تخصصی و عمومی توصیه می‌شود. به علاوه استفاده از روش‌های بیزی با توجه به وجود اطلاعات قبلی در وزنی کردن دروس هر بخش، و برآورد پارامترها با الگوریتم EM[1] توصیه می‌شود.

1. Expectation Maximization

منابع

- Bond, TG and Fox, CM (2007). *Applying the Rasch model: Fundamental measurement in the human sciences (2nd Ed)*, Mahwah, NJ, USA: Lawrence Erlbaum.
- Childs, R, Eligie, S, Gadalla, T, Traub, R and Jaciw, A (2004). IRT-linked standard errors of weighted composites, *Practical Assessment. Research and Evaluation*, *9*, (13).
- Cronbach, L, Glesser, G, Nanda, H and Rajaratnam, N (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*, Chichester: Wiley.
- Feldt, L and Brennan, R (1989). Reliability, Educational Measurement (*3*rd Edition, R Linn Ed), *The American Council on Education*, MacMillan, pp. 105-146.
- Gill, T and Bramley, T (2008). Using simulated data to model the effect of inter-marker correlation on classification, *Research Matters: A Cambridge Assessment Publication*, 5, pp. *29-36*.
- Hambleton R and Swaminathan, H. (1983). *Item response theory: Principles and applications*, the Netherlands: Kluwer-Nijoff.
- Hambleton, R, Swaminathan, H, and Rogers, J (1991). *Fundamentals of Item Response Theory,* Newbury Park, Ca, 12, pp. *177- 184*.
- Runder, L (2001). Informed test component weighting, *Educational Measurement: Issues and Practice*, 20, pp. 16-19.
- Wang, M and Stanley, J (1970). Differential weighting: A review of methods and empirical studies, *Review of Educational Research*, 40, pp. 663-705.
- Webb, N; Shavelson, R and Haertel, E. (2007). Reliability coefficient and generalizability theory, *Handbooks of Statistics 26: Psychometrics* (C Rao and S Sinharay Eds), pp. 81-120.
- Wright, B and Masters G (1982). *Rating scale analysis, Rasch Measurement*, Chicago, IL, USA: MESA Press.
- Wu, M and Adams, R (2006). Modelling mathematics problem solving item responses using a multidimensional IRT model, *Mathematics Education Research Journal*, 18, pp. 93-113.