Educational Measurement and Evaluation Studies

National Organization for Educational Testing

Print ISSN: 2476-2865 Online ISSN: 2783-0942

Introduction and application of Dichotomous Moken Scale Analysis (MSA) in educational and psychological measurements

Jalil Younesi[۱], Hengameh Aminaee[۲] , Reyhaneh Rezazadeh[۳]

۱. Associate Professor, Psychology Assessment and Measurement, Allameh Tabataba'i University, Tehran, Iran. jalilyounesi@gmail.com

۲. PhD Candidate in Psychology Assessment and Measurement, Allameh Tabatabai University, Tehran, Iran. h_aminaee@atu.ac.ir

۳. PhD Candidate in Psychology Assessment and Measurement, Allameh Tabatabai University,Tehran, Iran. rey_rezazadeh@atu.ac.ir

Article Info

Article Type:
Research Article

Received:
۲۰۲۱-۰۵-۰۱

Received in revised form:
۲۰۲۲-۰٤-۱٤

Accepted:
۲۰۲۲-۰٤-۱۹

Published online:
۲۰۲۱-۱۲-۲۲

ABSTRACT

Objective: This study aims to introduce this scale for Dichotomous Responses and shows the application of it in the field of educational measurements/assessments by providing evidence of fitting Mokken in study the level of knowledge of the fifth- and sixth-year students of Kerman dental school (academic year ۹۵-۹٦) about blood pressure (BP)measurement technique according the guideline of the American Heart Association.

Methods: Mokken scale analysis (MSA) is a probabilistic-nonparametric approach to Item Response Theory (IRT) that can be used to evaluate fundamental measurement properties in educational and psychological measurements with less strict assumptions than parametric IRT models.

Results : The results of secondary data analysis of the test included ۱٤ items and all of ۱۳۱ students, showed (using R software), that, the scalability coefficients of each item and the scalability coefficients of the item pairs were positive with low standard error. The overall scalability coefficient was ۰,٤۹۰ (medium), and omission of item ۵ increased the overall scalability coefficient to ۰,۵٤ (strong/high). Therefore, the results under the monotone homogeneity Model (MH) showed no violations of monotonicity and one-dimension was obtained. Invariant item-ordering Results from IIO analyses showed the number of violations from the assumption of non-intersecting response functions, bottom-up method for selecting items confirmed the deletion of item ۵. Obtaining the H^T scalability coefficient of ۰,٤۵, was an evidence of fitting double monotonicity (DM) model. Four reliability coefficients calculated for the set of ۱۳ items ranged from ۰,۸٦ to ۰,۹۱.

Conclusion: This study showed that this scale can be used to rank both individuals and items.

Keywords: non-parametric item- response theory, Moken scale analysis, monotone homogeneity (MH), double monotonicity model (DM).

Cite this article: Younesi, Jalil. Aminaee, Hengameh. Rezazadeh, Reyhaneh. (۲۰۲۱). Introduction and application of Dichotomous Moken Scale Analysis (MSA) in educational and psychological measurements. *Educational Measurement and Evaluation Studies*, ۱۱ (۳۳): ۲۷ pages. DOI:

© The Author(s).

Publisher: National Organization of Educational Testing (NOET)

معرفی و کاربرد مقیاس دوارزشی موکن در اندازه‌گیری های آموزشی و روانشناختی

جلیل یونسی[۱] ، هنگامه امینائی[۲] ، ریحانه رضازاده[۳]

۱. دانشیار سنجش و اندازه‌گیری دانشگاه علامه طباطبائی، نویسنده مسئول (younesi@atu.ac.ir)،

۲ . دانشجوی دکتری سنجش و اندازه گیری دانشگاه علامه طباطبائی،

۳ . دانشجوی دکتری سنجش و اندازه گیری دانشگاه علامه طباطبائی،

اطلاعات مقاله

نوع مقاله:
مقاله پژوهشی

دریافت:
۱۴۰۰/۰۲/۱۱
اصلاح:
۱۴۰۱/۰۱/۲۵
پذیرش:
۱۴۰۱/۰۱/۳۰
انتشار:
۱۴۰۰/۱۰/۰۱

چکیده

هدف: این مقاله ضمن معرفی مقیاس موکن[۱](MSA) ، کاربرد آن در زمینه ارزیابی آموزشی را با ارائه شواهد برازش مدل‌های موکن در مطالعه سنجش میزان آگاهی از تکنیک صحیح اندازه‌گیری فشارخون دانشجویان دندانپزشکی کرمان (سال تحصیلی ۹۵-۹۶) بر اساس دستورالعمل انجمن قلب آمریکا نشان می‌دهد.

روش پژوهش: تحلیل مقیاس موکن، رویکردی غیرپارامتری از نظریه سوال-پاسخ است که می‌تواند در ارزیابی و اندازه‌گیری‌های آموزشی و روان‌شناختی با فرضیاتی سهل‌گیرانه‌تر از مدل‌های پارامتری نظریه سوال- پاسخ استفاده شود.

یافته‌ها: در نتایج تحلیل ثانویه داده‌ها بر روی آزمونی ۱۴ سوالی که به شیوهٔ سرشماری از ۱۳۸ نفر دانشجوی سال پنجم و ششم دندانپزشکی دانشگاه کرمان به‌عمل‌آمده بود (با استفاده از نرم‌افزار R)، ضرایب مقیاس‌پذیری تک‌تک سوال‌ها و ضرایب مقیاس‌پذیری جفت سوال‌ها، مثبت با خطای معیار کم به دست آمد. ضریب مقیاس‌پذیری کل ۰/۴۹۰ (در حد متوسط) بود که حذف سوال ۵، شاخص مقیاس‌پذیری کل را به ۰/۵۴ (در حد قوی) افزایش داد. لذا نتایج تحت مدل همگن یکنواخت[۲] (MH) موکن نشان داد که فرض یکنواختی برای آزمون وجود دارد و تک‌بعدی بودن حاصل‌شده است. فرض رتبه‌بندی نامتغیر سوال‌ها از طریق تعداد دفعات نقض مفروضه عدم تقاطع توابع پاسخ به سوال، مورد بررسی قرار گرفت و روش انتخاب سوال به روش پس‌رو حذف سوال ۵ را تأیید کرد. به دست آمدن ضریب مقیاس‌پذیری[۳] H^T به میزان ۰/۴۵، نشان داد که شواهد برازش مدل یکنواختی مضاعف[۴] (DM) وجود دارد. چهار ضریب پایایی به‌دست‌آمده برای مجموعه ۱۳ سوالی از ۰/۸۶ تا ۰/۹۱ متغیر بودند.

نتیجه‌گیری: شواهد نشان داد که از این مقیاس می‌توان هم برای رتبه‌بندی افراد و هم برای رتبه‌بندی سوال‌ها استفاده کرد.

واژه‌های کلیدی: نظریه سوال- پاسخ غیرپارامتری، تحلیل مقیاس موکن، مدل همگنی یکنواخت، مدل یکنواختی مضاعف.

استناد: یونسی، جلیل؛ امینائی، هنگامه؛ رضازاده، ریحانه (۱۴۰۰). معرفی و کاربرد مقیاس دوارزشی موکن در اندازه‌گیری های آموزشی و روانشناختی. *مطالعات اندازه‌گیری و ارزشیابی آموزشی*، ۱۱ (شماره۳۳)، ۲۷صفحه. DOI:

حق‌مؤلف © نویسندگان

ناشر: سازمان سنجش آموزش کشور

۱. Mokken scale analysis
۲. Monotone Homogeneity
۳ . Scalability
۴. Double Monotonicity

مقدمه (مبانی نظری و پیشینه پژوهش)

استفاده از آزمون‌ها و پرسشنامه‌ها در آموزش و روانشناسی بسیار رایج است. درواقع، آن‌ها بخشی جدایی‌ناپذیر از اقدامات روزمره مربیان، معلمان، پزشکان، دانشجویان یا فارغ‌التحصیلان در این زمینه‌ها هستند. تصمیمات و تشخیص‌های بالینی و همچنین یافته‌های تحقیقات به‌شدت به کیفیت اندازه‌گیری‌های مورداستفاده بستگی دارد. علاوه بر این، در آزمون‌های سرنوشت‌ساز که در آن داوطلبان بر اساس نتایج آزمون به شغل یا تحصیل دسترسی پیدا می‌کنند، دقت رتبه‌بندی داوطلبین از اهمیت بالایی برخوردار است. بنابراین تحقیق درباره رواسازی و اطمینان از اینکه آزمون‌ها به‌طور دقیق سطوح سازه مربوطه را منعکس می‌کنند ضروری است (بقایی، ۲۰۲۱).

در اندازه‌گیری‌های آموزشی، مدل‌های سوال- پاسخی[۱] که بیشتر بکار برده می‌شوند، بر اساس یک فرمول پارامتری ساخته می‌شوند که در آن، یک تابع جبری خاص، شکل تابع سوال- پاسخ[۲](IRF) را مشخص می‌کند و با استفاده از تبدیل‌هایی منجر به اندازه‌گیری در یک مقیاس با فاصله برابر می‌شود. درحالی‌که می‌توان خصوصیات اندازه‌گیری را با استفاده از یک روش غیرپارامتری نظریه سوال-پاسخ مانند تحلیل مقیاس موکن بررسی کرد (ویند[۳]، ۲۰۱۷).

موکن از دو مدل IRT غیرپارامتری تشکیل‌شده است که فرض می‌کنند رابطه بین θ و احتمال پاسخ صحیح توسط محدودیت‌های ترتیب کنترل می‌شود؛ یعنی برای دو نفر m و n با $\theta_m < \theta_n$ داریم:

$$P(X_i = ۱ \square \theta_m) \leq P(X_i = ۱ | \theta_n)$$

و این احتمالِ پاسخ‌های صحیح، تابع سوال-پاسخ را تشکیل می‌دهند. احتمال اینکه فرد m با مقدار θ کمتر به‌درستی به سوال i پاسخ دهد باید همیشه کوچک‌تر (مساوی) شخص n با مقدار θ بالاتر باشد. معادله بالا بیان می‌کند که IRF تابعی غیر کاهشی از θ است و این تنها محدودیت در IRF است؛ بنابراین در مدل‌های سوال-پاسخ غیرپارامتری، IRFها می‌توانند اشکال مختلفی داشته و بسیار نامنظم باشند. تا زمانی که رابطه بین θ و احتمال پاسخ صحیح غیرکاهشی باشد، هر شکلی برای IRF مجاز است (سیجتسما[۴]، مولنار[۵]، ۲۰۰۲).

شکل ۱، نمونه‌ای از IRF را نشان می‌دهد. در مدل‌های پارامتری IRT مانند مدل راش[۶]، شکل IRF با استفاده از یک تابع (مثلاً تابع لجستیک) مشخص می‌شود.

بیش از ۵۰ سال پیش، لوییز گاتمن[۷] تکنیک اسکلوگرام[۸] خود را معرفی کرد. با این مدل جبرگرایانه[۹] می‌توان افراد را به‌صورت «خطی» (با توجه به توانایی‌های آن‌ها) و سوال‌ها را (بر اساس دشواری آن‌ها) رتبه‌بندی کرد. از آن زمان به بعد، مدل گاتمن حداقل در دو جهت یکی از نظر آمار پارامتری بر اساس مدل راش (۱۹۶۰) و دیگری توسط مدل همگنی یکنواخت موکن (۱۹۷۱) خانواده‌ای از مدل‌های احتمالی «خطی» تعمیم داده‌شده است (آنلو[۱۰]، ۲۰۰۷).

در تجزیه و تحلیل داده‌ها بر اساس نظریه کلاسیک آزمون[۱۱]، تأکید بر سوال‌ها و آماره‌های نمونه‌ای است که برحسب نمرات خام محاسبه می‌شوند و شاید شناخته‌شده‌ترین این آماره‌ها شاخص همسانی درونی کودر-ریچادسون۲۰[۱۲] (KR۲۰) باشد؛ اما در نظریه صفت پنهان بر برآورد پارامتر سوال‌ها و افراد و واریانس برآوردگرها که تبدیل غیرخطی از نمرات خام هستند تأکید می‌شود (اندریچ[۲]، ۱۹۸۲).

---

۱. Item Response Models
۲. Item Response Function (IRF)
۳. Wind
۴. Sijtsma
۵. Molenaar
۶. Rasch
۷. Louis Guttman
۸. Scalogram
۹. Deterministic
۱۰. Ünlü
۱۱. Classical test theory (CCT)

همسانی درونی این است که سوال‌های یک آزمون همگی «یک‌چیز» را منعکس کنند. «چیز»، همان‌طور که توسط کرونباخ[۳] (۱۹۵۱) و لومزدن[۴] (۱۹۵۷) توضیح داده شده است، لزوماً به معنای یک عامل یا ویژگی نیست. هر سوال ممکن است شامل بیش از یک ویژگی یا ترکیبی از ویژگی‌ها باشد، اما اگر همه سوال‌ها به نسبت یکسان دارای این ویژگی‌ها باشند، آنگاه به نظر می‌رسد که آزمون ازنظر درونی همسان است. ایده همسانی درونی در نظریه صفت پنهان، ایده تک‌بعدی بودن است، به این معنا که صفت یک فرد ممکن است با یک مقدار واحد در یک پیوستار پنهان نشان داده شود. تک‌بعدی بودن در نظریه صفت پنهان، معادل همسانی درونی در نظریه کلاسیک آزمون، امری نسبی است و نه مطلق.

از دیدگاه عملیاتی مقیاس گاتمن ازنظر شواهد تک‌بعدی بودن، ایده‌آل تلقی می‌شود. ارتباط آن با مفاهیم همسانی درونی نظریه کلاسیک این است که هرچه پاسخ‌ها به این ایده‌آل نزدیک‌تر باشند، ارزش شاخص KR۲۰ بیشتر است. دستیابی به ایده‌آل مقیاس گاتمن در آزمون واقعی دشوار است و مانع اصلی، الزاماتی است که پاسخ‌های یک فرد به یک سوال را به شیوه‌ی معین محدود می‌کند؛ بنابراین درصورتی‌که پراکندگی دشواری سوال‌ها زیاد باشد و نیز هیچ دو سوالی در مقیاس نزدیک به هم نباشند، تحقق مقیاس گاتمن افزایش می‌یابد. همچنین اگر دامنه توانایی‌های افراد نسبتاً زیاد باشد و حتی‌الامکان طیف گسترده‌ای از دشواری سوال‌ها را پوشش دهد، کمک‌کننده است. میزان انطباق مجموعه‌ای از پاسخ‌ها با مقیاس گاتمن را می‌توان از شاخص‌های مختلف قابلیت بازسازی[۵] نشان داد (گاتمن، ۱۹۴۷).

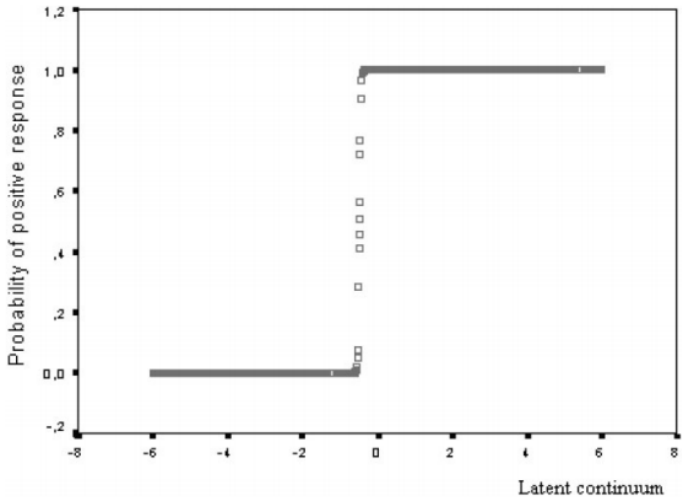
در مدل‌های IRT احتمال پاسخ صحیح به یک سوال دوارزشی به یک یا چند پارامتر فرد و یک یا چند پارامتر سوال بستگی دارد. یک مدل بسیار محبوبIRT، مدل لجستیک یک پارامتری (مدل PL۱) است که به نام مخترع آن، آماردان دانمارکی، جورج راش به‌عنوان مدل راش نیز شناخته می‌شود. در این مدل احتمال پاسخ صحیح به یک سوال به‌صورت زیر تعریف می‌شود (که θ_s پارامتر شخص s است، δ_i پارامتر سوال i است و α را می‌توان به‌عنوان یک ثابت برای یک مقیاس خاص در نظر گرفت):

$$P(X_{is} = ۱|\theta_s, \delta_i) = \frac{e^{\alpha(\theta_s - \delta_i)}}{۱ + e^{\alpha(\theta_s - \delta_i)}}$$

همان‌طور که در شکل ۱ نشان داده شده است، برای$\alpha = +\infty$ ، مدل به مقیاس گاتمن کاهش می‌یابد که در آن اگر$\theta_s > \delta_i$ ، احتمال پاسخ صحیح به ۱ نزدیک می‌شود و اگر $\theta_s < \delta_i$، احتمال پاسخ صحیح به ۰ نزدیک می‌شود و تابع سوال پاسخ که احتمال پاسخ صحیح به سوال i را توسط آزمودنی‌ها با پارامترهای مختلفِ شخص را نشان می‌دهد، یک تابع پلکانی[۶] به نظر می‌رسد(ون شوور[۷]،۲۰۰۳).

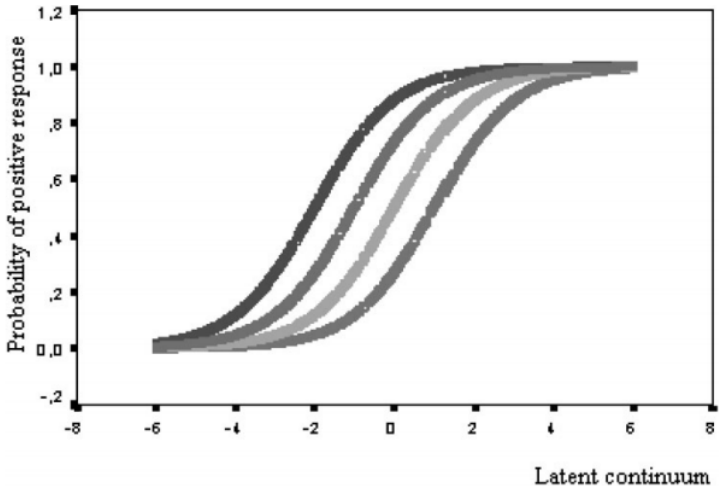
---

۱. Kuder – Richardson۲۰
۲. Andrich
۳. Cronbach
۴. Lumsden
۵. Reproducibility
۶. Step function
۷. Van Schuur



شکل ۱. تابع پاسخ پلکانی

برای مقادیر ثابت α (به‌عنوان‌مثال ۱= α)، همان‌طور که شکل ۲ نشان می‌دهد، IRF یک سوال خاص یک تابع لجستیک (S شکل) است و IRF ها برای سوال‌های دیگرِ یک مقیاس، با فرض برابر بودن α هایشان، همگی موازی هستند (ون شوور،۲۰۰۳).



شکل ۲. توابع پاسخ موازی

مقیاس موکن یک رویکرد احتمالی غیرپارامتری از نظریه سوال–پاسخ است که چارچوبی منظم برای ارزیابی کیفیت اندازه‌گیری‌ها فراهم می‌کند. مدل‌های مبتنی بر مقیاس موکن، غیرپارامتری در نظر گرفته می‌شوند و رابطه بین متغیر پنهان و احتمال پاسخ (تابع سوال–پاسخ) مادامی‌که مفروضه‌های اصلی ترتیب برآورده شود، لازم نیست با شکل خاصی مطابقت کنند. اگر یک مدل IRT را برای ساختن آزمونی به کار ببریم و اندازه‌گیری پاسخ‌دهندگان بر اساس یک مقیاس ترتیبی برای کارکردهای پیش‌بینی‌شده کافی باشد، استفاده از مدل‌های پارامتری IRT برای چنین اهداف پیش‌بینی‌شده‌ای می‌تواند بی‌جهت ایجاد محدودیت کند. برازاندن یک مدل پارامتری با IRF های لجستیک بر داده‌ها می‌تواند بیشتر سوال‌ها را به این دلیل که نمی‌توانند ارتباط کارکردی با توانایی را به‌وسیله یک تابع لجستیک الگوسازی کنند، رد کند، درحالی‌که گنجاندن سوال‌ها با IRF های غیرکاهشی، دقت اندازه‌گیری را افزایش می‌دهد (موکن، ۱۹۷۱).

مقیاس موکن به محققان کمک می‌کند سوال‌هایی را که بر طبق مدل‌های محدودکننده IRT پارامتری باید حذف کرد، حفظ کرد. همچنین MSA مجموعه‌ای از ابزارهای اکتشافی را برای تجزیه‌وتحلیل ابعاد فراهم می‌کند که در مدل‌های پارامتری IRT امکان‌پذیر نیست (امانز[۱]، سیجتسما و پدرسن[۲]، ۲۰۱۲). بعضی از متخصصان آزمون‌سازی استدلال می‌کنند که در عمل اکثر IRF ها، لجستیک یا چیزی نزدیک به آن هستند. مخالفان این بحث می‌گویند که استفاده از مدل تک‌پارامتری لجستیک باعث رد همه سوال‌هایی می‌شود که شیب IRF های آن‌ها متفاوت با مدل تک‌پارامتری است. استفاده از مدل‌های دوپارامتری هم باعث رد سوال‌هایی می‌شود که مجانب پایین IRF های آن‌ها بزرگ‌تر از صفر است (به‌عنوان نمونه سوال‌ها چندگزینه‌ای) و استفاده از مدل سه‌پارامتری لجستیک باعث می‌شود که سوال‌های که مجانب بالای آن‌ها کمتر از یک است و همچنین سوال‌های که IRF هایی با شکل نامنظم دارند و یک یا بیش از یک خمیدگی تیز دارند از آزمون یا پرسشنامه حذف شوند(سیجتسما، مولنار، ۲۰۰۲). هدف اصلی مدل موکن، رواسازی مقیاس رتبه‌ای متغیر پنهان است. برای سوال‌هایی که پیش‌فرض‌های مدل موکن را برآورده می‌کنند، می‌توان از مجموع پاسخ‌های موجود به سوال‌ها برای رتبه‌بندی پاسخ‌دهندگان در ویژگی پنهان استفاده کرد (همکر[۳]، سیجتسما، مولنار و جانکر[۴]، ۱۹۹۷). در مقایسه با مدل‌های IRT پارامتری، مدل موکن با توجه به رابطه بین صفت پنهان و پاسخ به سوال‌ها به مفروضه‌های کمتری نیاز دارد؛ بنابراین امکان نگه‌داشتن سوال‌های مهم‌تر را فراهم می‌کند. درنتیجه ترتیب افراد دقیق‌تر است (سیجتسما و مولنار، ۲۰۰۲). بسیاری از محققین به‌ویژه چرنیسنکو[۵]، استارک[۶]، چان[۷]، دراسگو[۸] و ویلیامز[۹]، ۲۰۰۱؛ و همچنین میجر[۱۰] و بنکر[۱۱]، ۲۰۰۴، میجر، تندرو[۱۲] و ندرز[۱۳]، ۲۰۱۴ و رایس[۱۴] و والر[۱۵]، ۲۰۰۹ اشاره کرده‌اند که رویکردIRT غیرپارامتری به‌طورکلی و به‌طور خاص مقیاس موکن در زمینه‌هایی که در آن فرآیندهای پاسخ زیربنایی (مانند متغیرهای عاطفی[۱۶]) به‌خوبی درک نمی‌شوند، مفید است. اگرچه مقیاس موکن به‌طور گسترده در انواع حوزه‌های عاطفی مورداستفاده قرارگرفته است، استفاده از این روش در زمینه ارزیابی‌های آموزشی کمتر متداول است. در آزمون‌های پیشرفت تحصیلی نیز فرایندهای پاسخگویی آزمون‌شوندگان به‌خوبی موردمطالعه قرار نگرفته است و مقیاس موکن چارچوبی مفید برای کاوش در خصوصیات اساسی اندازه‌گیری این ارزیابی‌ها ازجمله درجهٔ تغییرناپذیری دانش‌آموزان و ترتیب سوال‌ها، فراهم می‌کند. به‌عنوان‌مثال سنجش پیشرفت تحصیلی فرایندی پیچیده است که از داوران خواسته می‌شود تا درک خود از پیشرفت دانشجو را در یک مقیاس رتبه‌بندی منعکس کنند. داوری با استفاده از متغیرهایی مانند کارهای دانشجویی، مقالات، مقیاس‌های رتبه‌بندی و ویژگی‌های رقابتی فرد صورت می‌پذیرد. به همین ترتیب فرآیندهای پاسخگویی برای سوال‌ها چندانتخابی[۱۷] شامل تعامل بین مکان فرد در یک متغیر پنهان، قالب‌های سوال، محتوای مورد ارزیابی و سایر ویژگی‌های فرد است. در این حالت به‌کاربردن مقیاس موکن در ارزیابی‌های آموزشی مناسب به

---

۱. Emons
۲. Pedersen
۳. Hemker
۴. Junker
۵. Chernyshenko
۶. Stark
۷. Chan
۸. Drasgow
۹. Williams
۱۰. Meijer
۱۱. Baneke
۱۲. Tendeiro
۱۳. Wanders
۱۴. Reise
۱۵. Waller
۱۶. Affective
۱۷. Multiple choice

نظر می‌رسد. درواقع هنگامی‌که محققان در سنجش‌های تشخیصی و چه در سنجش‌های تکوینی علاقه‌مندند که میزان مطابقت یک ارزیابی با خصوصیات اساسی اندازه‌گیری شده را بررسی یا پیگیری کنند، استفاده از مقیاس موکن مفید است. همچنین در زمینه‌هایی که در آن اطلاعاتی در مورد کیفیت اندازه‌گیری و مرتب کردن افراد یا سوال‌ها نیاز است، اما حجم نمونه برای دستیابی به برآوردهای باثبات بر اساس مدل‌های پارامتری نظریه سوال–پاسخ پارامتری کافی نیست، این رویکرد کاربرد دارد (وبند، ۲۰۱۷).

تحقیقات بر پایه مدل‌های مقیاس موکن در مواردی که شواهد پیروی از مفروضه‌های مدل‌های آن مهیاست، از تفسیر نمرات کل (جمع نمرات) به‌منظور مرتب کردن معنادار افراد و سوال‌ها روی متغیر پنهان و بدون هرگونه تبدیل پارامتری، حمایت می‌کنند. به‌علاوه در مواردی که نتیجه‌گیری مطلوب از یک روش ارزیابی به سطح مقیاس فاصله‌ای نیاز نداشته باشد (سنجش انطباقی با رایانه یا بعضی از روش‌های همترازسازی)، اما بتوان از اطلاعات رتبه‌ای مرتبط با دانش‌آموزان و سوال‌ها مطلع بود؛ مقیاس موکن مناسب است (سیجتسما و میجر، ۲۰۰۷).

تحلیل مقیاس موکن مجموعه‌ای از ابزارهای آماری را برای ساخت مقیاس، اندازه‌گیری توانایی افراد و بررسی سوال‌ها در حوزه‌های شخصیتی، شناختی، کیفیت زندگی مرتبط با سلامتی، جامعه‌شناسی، بازاریابی و زمینه‌های دیگر که ابزارهایی با مجموعه‌ای از سوال‌ها است را فراهم می‌کند. این روش از مفروضه‌های نسبتاً سهل‌گیرانه‌ای استفاده می‌کند که حکایت از ترتیب افراد در یک مقیاس با استفاده از نمرات کل افراد در مجموعه سوال‌ها دارد (سیجتسما و مولنار، ۲۰۰۲).

MSA همچنین در مواردی که نتیجه‌گیری مطلوب از یک روش ارزیابی به سطح مقیاس فاصله‌ای نیاز نداشته باشد اما بتوان از اطلاعات رتبه‌ای مرتبط با آزمون‌شوندگان و سوال‌ها مطلع بود؛ مناسب است(وبند، ۲۰۱۷). هنگامی‌که سطح مقیاس فاصله‌ای موردنیاز است(سنجش انطباقی با رایانه یا بعضی از روش‌های همترازسازی) رویه‌های مبتنی بر مقیاس موکن می‌تواند یک مرور اولیه اکتشافی در مورد درجه‌ای که طبق آن روش اندازه‌گیری به الزامات اصلی اندازه‌گیری پایبند است و می‌تواند به‌عنوان لنزی عمل کند تا بتوان داده‌ها را با استفاده از نتایج عددی و نمایشگرهای گرافیکی موردبررسی قرارداد، فراهم کند(سیجتسما و میجر، ۲۰۰۷). MSA به‌سرعت در بین محققان کاربردی محبوبیت پیداکرده است؛ جستجو در ادبیات نشان می‌دهد که از سال ۲۰۱۵ پژوهش‌های بسیاری با استفاده از این مقیاس انجام‌گرفته، اما تاکنون پژوهش‌های چندانی با استفاده از این مقیاس در کشور ما صورت نگرفته است. هدف اصلی این مقاله، ارائه مقدمه‌ای مختصر از ویژگی‌های اصلی رویکرد مقیاس موکن و کاربردهای آن در ارزیابی‌های آموزشی و روان‌شناختی است.

مقیاس بندی مدل‌های موکن برای پاسخ‌های دوارزشی

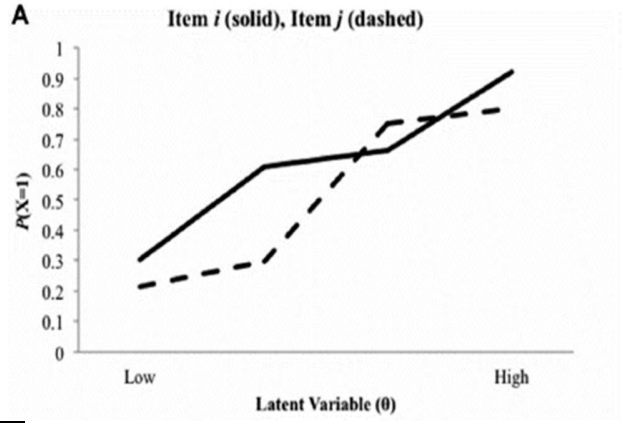
ریشه MSA را می‌توان در مقیاس گاتمن جستجو کرد. این مقیاس توسط لوئیس گاتمن (۱۹۵۰) برای سوال‌ها دوارزشی با هدف استخراج یک بعدِ واحد از داده‌ها و قرار دادن افراد و سوال‌ها در این بعد با مقدار عددی ایجاد شد(وبند، ۲۰۱۷). جدول ۱. نمونه‌ای از مقیاس کامل گاتمن را برای یک آزمون زبان نشان می‌دهد. سوال‌ها ۱ تا ۵ از ساده‌ترین به سخت‌ترین و آزمون‌شوندگان A تا E از تواناترین به کم‌توان‌ترین مرتب‌شده‌اند. انتظار داریم فردی که به سوال ۲ پاسخ صحیح داده است به سوال ۱ نیز پاسخ صحیح داده باشد. به‌طور مثال فردی که نحوه تشکیل سوال‌ها را در زمان گذشته یاد گرفته است (سوال ۲) باید زمان حال ساده را نیز یاد گرفته باشد (سوال ۱). زبان‌آموزی که به حد معینی از توانایی گرامری رسیده باشد می‌تواند به تمام سوال‌های زیر این سطح به‌درستی پاسخ دهد و هیچ‌یک از سوال‌های بالای آن را نتواند پاسخ دهد. به‌وسیله نمره کل، بازتولید بردار پاسخ یک فرد نیز باید امکان‌پذیر باشد. نمره کل ۲ به معنی الگوی پاسخ ۱۱۰۰۰ و نمره کل ۴ به معنی الگوی پاسخ ۱۱۱۱۰ است. اگر مقیاس گاتمن برقرار باشد، گواه این است که سوال‌ها با یک مجموعه تک‌بعدی مطابقت دارند و افراد در یک پیوستار مرتب‌شده‌اند. پاسخ‌های غیرمنتظره به‌عنوان خطاهای گاتمن نامیده می‌شوند (بقایی، ۲۰۲۱).

جدول ۱. نمونه‌ای از مقیاس کامل گاتمن

سوال‌ها					
۵	٤	۳	۲	۱	آزمون‌شوندگان
۱	۱	۱	۱	۱	A
۰	۱	۱	۱	۱	B
۰	۰	۱	۱	۱	C
۰	۰	۰	۱	۱	D
۰	۰	۰	۰	۱	E

در ارائه اصلی مقیاس موکن (۱۹۷۱) برای ارزیابی کیفیت مقیاس‌های ساخته‌شده از سوال‌های دوارزشی (۱ یا ۰ $X = ۰$)، دو مدل ارائه شد. برای اینکه پاسخ‌های دوارزشی برای مقیاس موکن مناسب باشند، نمره ۱ باید موقعیت مکانی بالاتر در متغیر پنهان (توانایی بالاتر درآزمون‌های پیشرفت تحصیلی) نسبت به نمره ۰ را نشان دهد. مدل اول، مدل همگنی یکنواخت است که از بین دو مدل اصلی مقیاس موکن، عمومی‌تر است (وبند، ۲۰۱۷). مدل همگن یکنواخت مبتنی بر سه فرض اساسی به شرح زیر است: (۱) یکنواختی[۱]: با افزایش مکان در متغیر پنهان، احتمال پاسخ صحیح ($X = ۱$) کاهش نمی‌یابد. (۲) تک‌بعدی بودن[۲]: پاسخ‌های سوال، نشانگر شواهدی از تنها یک متغیر پنهان است. (۳) استقلال موضعی[۳]: پاسخ به یک سوال پس از کنترل متغیر پنهان، تحت تأثیر پاسخ به سوال‌ها دیگر نیست (سیجتسما و وندرارک، ۲۰۱۷).

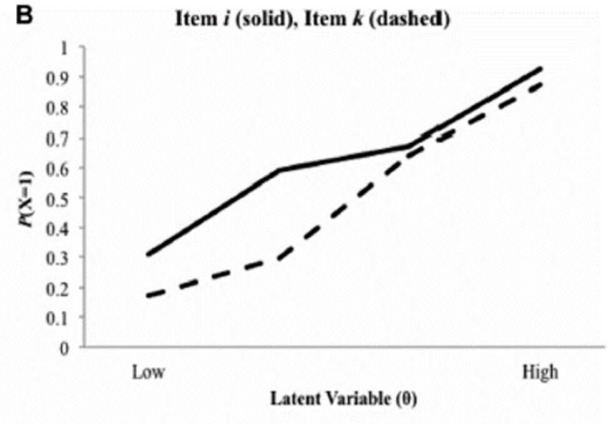
چندین ویژگی در مورد مدل همگن یکنواخت قابل‌توجه است: ۱) مدل همگن یکنواخت تابع پاسخ سوال را به‌جز فرض یکنواختی بر پیش‌فرض خاصی محدود نمی‌کند. درنتیجه تابع پاسخ سوال‌هایی که به مدل همگن یکنواخت پایبند هستند ممکن است اشکال مختلفی به خود بگیرد که لزوماً با شکل اجایو لجستیک که معمولاً با مدل‌های پارامتریک نظریه سوال–پاسخ همراه است، مطابقت ندارند (وبند، ۲۰۱۷). شکل ۳ نشان‌دهنده یک جفت تابع پاسخ سوال برای دو سوال است که با مفروضه‌های مدل‌های همگن یکنواخت مطابقت دارند. محور Y، احتمال پاسخ صحیح (تابع پاسخ سوال) و محور X متغیر پنهان (θ) را نشان می‌دهد. اگرچه توابع پاسخ سوال‌های i و j متقاطع هستند، اما آن‌ها به الزامات مدل همگن یکنواخت پایبند هستند، زیرا در اثر افزایش مکان در متغیر پنهان کاهش نمی‌یابند (سیجتسما و وندرارک، ۲۰۱۷).



۱. Monotoniciy
۲. Unidimensionality
۳. Local independence

شکل ۳. توابع پاسخ غیرکاهشی و متقاطع

هنگامی‌که داده‌ها متناسب با مفروضه‌های مدل همگن یکنواخت هستند، ترتیب نسبی آزمون‌شوندگان در متغیر پنهان در تمام سوال‌ها، ثابت است. ازآنجاکه متناسب با مفروضه‌های، مدل همگن یکنواخت شواهدی برای ترتیب ثابت افراد در سوال‌ها فراهم می‌کند، این مدل می‌تواند مشابه مدل لجستیک دو پارامتری در نظریه سوال–پاسخ پارامتری باشد (ویند، ۲۰۱۷). ۲) مدل دوم موکن (۱۹۷۱) مدل یکنواختی مضاعف است که یک حالت خاص از مدل همگن یکنواخت به‌حساب می‌آید. مدل همگن مضاعف در سه فرض مدل همگن یکنواخت مشترک است، اما فرض چهارمی را هم شامل می‌شود: (۴) ترتیب‌بندی نامتغیر سوال‌ها یا نامتغیر بودن ترتیب‌بندی سوال‌ها[۱] (IIO): توابع پاسخ برای سوال‌های خاص با توابع پاسخ برای هر سوال دیگر تلاقی ندارند (سیجتسما و وندرارک، ۲۰۱۷). شکل ۴ نشانگر یک جفت تابع برای دو سوال دوارزشی است که با مفروضه‌های مدل همگن مضاعف مطابقت دارند زیرا توابع پاسخ سوال‌های i و k هر دو یکنواخت (غیر کاهشی در سرتاسر متغیر پنهان) و غیرمتقاطع هستند (ویند، ۲۰۱۷).



شکل٤. توابع پاسخ غیرکاهشی و غیرمتقاطع

نتیجه مهم این پیش‌فرض‌ها برای سوال‌های دوارزشی این است که وقتی داده‌ها با مفروضه‌های مدل همگن مضاعف متناسب هستند، سوال‌ها به روش یکسان در بین دانش‌آموزان مرتب می‌شوند؛ زیرا تناسب با مفروضه‌های مدل همگن مضاعف، شواهدی برای ترتیب بندی ثابت هم برای آزمون‌شوندگان و هم برای سوال‌ها فراهم می‌کند، این مدل به‌عنوان نسخه رتبه‌ای مدل راش دوارزشی یا مدل لجستیک یک پارامتری در نظریه سوال– پاسخ پارامتری توصیف‌شده است (انگلهارد[۲]، ۲۰۰۸).

مقیاس‌پذیری

موکن (۱۹۷۱) بسطهایی از ضرایب مقیاس‌پذیری که ابتدا توسط لووینگر[۳] (۱۹۴۸) ارائه‌شده است را برای توصیف درجه‌ای که هر سوال و مجموعه‌های کلی از سوال‌های مقیاس می‌توانند برای مرتب کردن افراد در یک سازه استفاده شوند را توسعه داد. شاخص‌های مقیاس‌پذیری خلاصه‌ای از تأثیر خطاهای گاتمن بر کیفیت یک روش اندازه‌گیری را ارائه می‌دهند. خطاهای کمتر گاتمن تفسیر نمرات کل افراد را به‌عنوان شاخص‌های مرتب کردن فرد در سازه تسهیل می‌کند. (ویند، ۲۰۱۷).

۱. Invariant item ordering (IIO)
۲. Engelhard
۳. Loevinger

موکن از سه ضریب مقیاس‌پذیری استفاده می‌کند. ضریب مقیاس‌پذیری سوال (H_i)، ضریب مقیاس‌پذیری جفت سوال (H_{ij}) و ضریب مقیاس‌پذیری کل (H^T). ضریب مقیاس‌پذیری با استفاده از نسبت فراوانی‌های خطای گاتمن مشاهده‌شده و مورد انتظار محاسبه می‌شوند. ضرایب مقیاس‌پذیری به‌عنوان یک منهای نسبت خطاهای گاتمن مشاهده‌شده (F_{ij}) و مورد انتظار (E_{ij}) برای هر سوال، برای جفت سوال‌ها و کل آزمون محاسبه می‌شود. ضریب مقیاس‌پذیری برای جفت سوال‌ها:

$$H_{ij} = ۱ - \frac{F_{ij}}{E_{ij}}$$

که در آن F_{ij} تعداد خطاهای گاتمن برای جفت سوال i و j است و E_{ij} تعداد مورد انتظار خطاهای گاتمن تحت استقلال حاشیه‌ای ($\frac{\text{سطر فراوانی} \times \text{ستون فراوانی}}{\text{کل فراوانی}}$) است. این روش معادل محاسبه کواریانس دو سوال است (سیجتسما و مولنار،۲۰۰۲). ضریب مقیاس‌پذیری جفت سوال (H_{ij})به‌عنوان نسبت کواریانس بین هر جفت از سوال‌های i و j حداکثر کواریانس ممکن آن‌ها با توجه به توزیع حاشیه‌ای نمرات دو سوال تعریف می‌شود و منعکس‌کنندهٔ سازگاری درونی هر جفت سوال است (موویج[۱]، ۲۰۱۲).

به همین ترتیب ضرایب مقیاس‌پذیری برای هریک از سوال‌ها H_i با استفاده از فرمول بالا با در نظر گرفتن تعداد خطاهای گاتمن مشاهده‌شده و مورد انتظار برای همه جفت سوال‌های مرتبط با سوال موردنظر محاسبه می‌شود. دو سوال i و j را طوری در نظر گرفته می‌شود که i سخت‌تر از j باشد. جدول ۲×۲ زیر نتایج احتمالی را نشان می‌دهد.

جدول ۲. نتایج احتمالی را برای دو سوال

	سوال j	
سوال i	۱	۰
۱	A(۱,۱)	B(۱,۰)
۰	C(۰,۱)	D(۰,۰)

سلول‌های A ، B ، C و D به ترتیب بیانگر تعداد پاسخ‌دهندگانی است که به هر دو سوال به‌درستی پاسخ داده‌اند، به سوال i درست پاسخ داده‌اند اما سوال j را اشتباه گفته‌اند، به سوال i اشتباه اما به سوال j صحیح پاسخ داده‌اند و آن‌هایی که به هر دو سوال اشتباه پاسخ داده‌اند. ازآنجایی‌که سوال j ساده‌تر از سوال i است، انتظار داریم سلول C بزرگ‌تر از سلول B باشد. ارزیابی این فراوانی‌ها نشان می‌دهد که آیا محدودیت‌های ترتیب برای سوال‌ها حفظ‌شده است یا خیر. این دقیقاً شبیه به بررسی این است که آیا خطاهای گاتمن رخ‌داده است یا خیر. هرچه تعداد خطاها کمتر باشد، پاسخ‌ها بیشتر از الگوی گاتمن پیروی می‌کنند و سلسله مراتبی که توسط سوال‌ها تشکیل می‌شود قوی‌تر است (بقایی، ۲۰۲۱).

ضریب مقیاس‌پذیری سوال (H_i) به‌صورت نسبت جمع کل کواریانس‌های جفتی با توجه به هر سوال i و مجموع حداکثر کواریانس‌های دوتایی این سوال بیان می‌شود. به‌طور خلاصه دقت تمییز سوال و قدرت رابطه بین سوال و مجموعه کل سوال‌ها است (امانز[۲] و همکاران، ۲۰۱۲). مقادیر بالاتر H_i نشان‌دهندهٔ قدرت تشخیص بالاتر است (موویج، ۲۰۱۲).

وقتی داده‌ها با مدل MH مطابقت دارند، مقادیر ضرایب مقیاس‌پذیری مثبت و از صفر تا یک متغیر هستند، مقدار یک نشان‌دهنده عدم وجود خطاهای گاتمن و مقادیر نزدیک به صفر نشان‌دهنده بسیاریِ خطاهای گاتمن است (وبند، ۲۰۱۷)؛ درنهایت ضریب مقیاس‌پذیری برای کل آزمون، H^T، با شمارش تعداد خطاهای گاتمن مشاهده‌شده و مورد انتظار برای کل داده‌ها با فرض حفظ یک ساختار کامل گاتمن محاسبه می‌شود. ضریب H^T برای کل مقیاس نقش محوری در مدل همگنی یکنواخت دارد و نشان می‌دهد که ساختار داده تا چه حد به الگوی گاتمن

۱. Mooij
۲. Emons

کامل نزدیک است (بقایی،۲۰۲۱). ضریب H^T یک شاخص دقیق از تک‌بعدی بودن و معنی‌دار بودن نمره کل برای مرتب کردن افراد است (سیجتسما و مولنار،۲۰۰۲). موکن (۱۹۷۱) اظهار داشت که «ما ترجیح می‌دهیم از ضریب همگنی H^T ، به‌عنوان معیار مقیاس‌پذیری به معنای همگنی یکنواخت استفاده کنیم. ضریب مقیاس‌پذیری ما به‌عنوان تنها معیار مقیاس‌پذیری ما خواهد بود».

اگر ضریب مقیاس‌پذیری کل بین ۰/۳ و ۰/۴ باشد ($0/3 \leq H^T < 0/4$) مقیاس ضعیف در نظر گرفته می‌شود، ضریب بین ۰/۴ تا ۰/۵، ($0/4 \leq H^T < 0/5$) مقیاس متوسط و اگر بزرگ‌تر از ۰/۵، ($H^T \geq 0/5$) باشد، قوی در نظر گرفته می‌شود. مقادیر H_{ij} باید بیشتر از صفر باشد و سرانجام اگر ضریب ($H_j < 0/3$)، سوال j باید بررسی یا حذف شود، اما اگر ($0/3 \leq H_j$) باشد، سوال می‌تواند برای تشکیل مقیاس موکن انتخاب شود (عبدالحمید[۱] و همکاران، ۲۰۲۰).

ضریب مقیاس‌پذیری کل H^T نسبت مجموع تمام کواریانس‌های جفتی و جمع کل حداکثر کواریانس‌های جفتی است و مقادیر بالاتر H^T نشان می‌دهد که از میانگین نمره کل می‌توان برای مرتب کردن افراد با دقت بالا استفاده کرد (موویج، ۲۰۱۲).

یکنواختی

همان‌طور که قبلاً توضیح داده شد، یکنواختی به این معنی است که با افزایش مکان افراد در متغیر پنهان، احتمال پاسخ صحیح یا باید افزایش یابد یا ثابت بماند. چندین روش برای ارزیابی یکنواختی ازجمله بررسی نمرات مانده گروهی[۲] وجود دارد. نمره مانده[۳] عبارت از امتیاز یا نمره خام کل بدون احتساب نمره سوال مورد بررسی است. هر نمره مانده باید حداقل تعداد اعضا ثابتی داشته باشد. سپس با استفاده از روش‌های گرافیکی و روش‌های آماری می‌توان یکنواختی را ارزیابی کرد. نمره مانده گروهی و نسبت پاسخ‌دهندگانی است که به یک سوال به‌درستی پاسخ داده‌اند (احتمال پاسخ صحیح) در محورهای x و y رسم می‌شوند. با افزایش نمره مانده، احتمال پاسخ صحیح باید افزایش یابد یا ثابت بماند. به این معنی که IRF ها باید نسبت به مقادیر فزاینده نمره مانده که مقادیر θ بالاتر را نشان می‌دهند، غیر کاهشی باشند (وینه، ۲۰۱۷).

آزمونی را با ۱۰ سوال در نظر بگیرید. فرض کنید می‌خواهیم یکنواختی را برای سوال ۱۰ ارزیابی کنیم. به‌استثنای سوال ۱۰، ۱۰ گروه نمره مانده وجود دارد، یعنی دانش‌آموزانی که در کل آزمون نمره ۰ گرفتند، آن‌هایی که نمره ۱ گرفتند، آن‌هایی که نمره ۲ ... و کسانی که نمره ۹ گرفتند. جدول ۳ نشان می‌دهد که در گروهی که در کل آزمون نمره ۰ کسب کرده‌اند، هیچ‌کس به سوال ۱۰ پاسخ صحیح نداده است. در گروهی که در کل آزمون نمره ۱ را کسب کرده‌اند، ۵ درصد به سوال ۱۰ پاسخ صحیح داده‌اند و ... (بقایی، ۲۰۲۱).

جدول ۳. نمرات باقی‌مانده گروهی و نسبت‌های مرتبط با آن‌ها

نمرات باقی‌مانده	۰	۱	۲	۳	٤	٥	٦	٧	٨	٩
نسبت پاسخ‌های صحیح	۰/۰	۰/۰٥	۰/۰۹	۰/۱٤	۰/۲٥	۰/٤۰	۰/٤۹	۰/٦٥	۰/۷۲	۰/۸٤

اگر گروه‌های باقی‌مانده کوچک باشند، نسبت پاسخ‌های صحیح، برآورد قابل‌اعتمادی از احتمال پاسخ صحیح نیست. اگر این اتفاق بیفتد، گروه‌های باقی‌مانده با گروه‌های مجاور ترکیب می‌شوند. اگر $N > ۵۰۰$ باشد، حداقل تعداد برای گروه نمرات باقی‌مانده $\frac{N}{۱۰}$ است. اگر $۲۵۰ < N < ۵۰۰$ باشد، حداقل تعداد افراد برای گروه‌های باقی‌مانده $\frac{N}{۵}$ است؛ اگر $N < ۲۵۰$ باشد، حداقل اندازه گروه‌های باقی‌مانده $\frac{N}{۳}$ است (سیجتسما و مولنار،۲۰۰۲). این فرمول‌ها تضمین می‌کنند که حداقل تعداد در گروه‌های باقی‌مانده ۵۰ باشد.

---

۱. Abdelhamid

۲. Restscore groups

۳. Restscore

رتبه‌بندی نامتغیر سوال‌ها[۱]

در آزمون‌های روان‌شناختی و آموزشی، سوال‌ها برای نشان دادن سطوح مختلف سازه نوشته می‌شوند. در یک آزمون گرامر انگلیسی، سوال شرطی نوع سوم[۲] باید دشوارتر از سوال‌های مربوط به مبحث مجهول[۳] باشد. فردی که به اولی به‌درستی پاسخ می‌دهد، نسبت به کسی که دومی را درست پاسخ می‌دهد، در گرامر انگلیسی مهارت بیشتری دارد. تفاوت در دشواری سوال‌ها، با میانگین سوال‌ها در جمعیت موردنظر اندازه‌گیری می‌شود. بااین‌حال، این ترتیب دشواری که از پاسخ‌های جمعیت، کمی سازی می‌شود ممکن است برای همه افراد صادق نباشد. باید به‌صورت تجربی بررسی شود که سوال شرطی نوع سوم درواقع دشوارتر از سوال مجهول برای همه امتحان شوندگان بوده است؛ یعنی ترتیب سوال‌ها به لحاظ دشواری، در بین افراد با سطوح متفاوت توانایی گرامری ثابت می‌ماند. سازندگان و کاربران آزمون باید بتوانند فرض کنند که فردی با نمره کل بالاتر به تمام سوال‌هایی که فردی با نمره کل پایین‌تر پاسخ داده است، به‌علاوه چند سوال دیگر به‌درستی پاسخ داده است. اگر مجموع نمرات، ترکیبی تصادفی از پاسخ‌های صحیح یا به عبارتی فاقد الگوهای قابل پیش‌بینی باشد، معنی‌دار بودن نمرات کل زیر سوال می‌رود (بقایی ۲۰۲۱).

اگر در یک پرسشنامه افسردگی یا اضطراب، IIO وجود داشته باشد، نشان می‌دهد که فردی با نمره کل بالاتر، تمام علائم فردی با نمره پایین‌تر و برخی علائم اضافی را دارد. علاوه بر این نمره کل را می‌توان هم به‌عنوان شاخصی از سطح صفت و هم به‌عنوان فهرستی از علائمی که بیمار از آن رنج می‌برد به‌طور معناداری تفسیر کرد؛ زیرا سلسله‌مراتب، علائم را جذب می‌کند (لیگتوت[۴] و همکاران، ۲۰۱۱).

علاوه بر این تحلیل الگوهای پاسخ افراد که به‌عنوان برازش فرد[۵] در ادبیات IRT شناخته می‌شود، تنها در صورتی منطقی است که IIO برقرار باشد. اگر الگوی پاسخ فردی به ۱۰ سوال که از ساده‌ترین تا سخت‌ترین مرتب‌شده‌اند، ۱۱۱۱۰۰۰۱۰۱ باشد، بدیهی است که پاسخ صحیح او به سوال‌های هشتم و دهم غیرمنتظره است و این به معنای عدم برازش فرد است زیرا این دو سوال سخت را در حالی به‌درستی پاسخ داده است کهسه سوال ساده‌تر را ازدست‌داده است. این نتیجه‌گیری تنها در صورتی معتبر است که بتوانیم فرض کنیم که ترتیب دشواری سوال‌ها برای همه امتحانشوندگان یکسان است، یعنی IIO برقرار است. اگر IIO پابرجا نباشد، ممکن است سوال‌های هشتم و دهم برای این آزمودنی خاص سخت نباشد و نتیجه‌گیری عدم برازش برای این فرد اشتباه باشد. وقتی یک سوال برای یک آزمودنی آسان و برای دیگری سخت باشد، تفسیر نمره کل مشکل خواهد بود. ازاین‌رو، IIO تفسیر نمره آزمون را تسهیل می‌کند (بقایی ۲۰۲۱). IIO را می‌توان با استفاده از چندین روش ازجمله روش مانده[۶]، تقسیم سوال[۷]، روش ماتریس نسبت[۸] و روش تقسیم نمرات مانده[۹] ارزیابی کرد (ون شوور،۲۰۱۱). در روش مانده‌ها، IIO برای همه جفت سوال‌ها به‌طور جداگانه ارزیابی می‌شود و عدم تلاقی IRF های آن‌ها بررسی می‌شود. مانده عبارت از نمره کل منهای نمره جفت سوال‌ها است. نمرات مانده بر روی محور x و احتمال پاسخ صحیح متناظرش (نسبت پاسخ‌های صحیح) در محور y رسم می‌شوند. به‌این‌ترتیب IRF ها ترسیم‌شده و عدم تقاطع آن‌ها بررسی می‌شود. اگر IRF ها ترتیب نسبی یکسانی را در محدوده نمرات مانده داشته باشند، یعنی ترتیب آن‌ها معکوس نشود، خاصیت IIO برای جفت سوال‌ها باقی می‌ماند. علاوه بر بررسی گرافیکی، آزمون فرضیه‌های آماری نیز وجود دارد (بقایی، ۲۰۲۱).

---

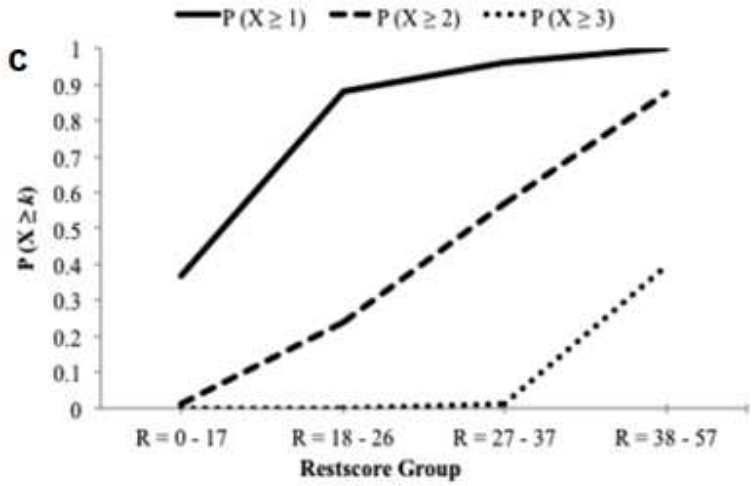
۱. Invariant item ordering
۲. The conditional type III
۳. Passive voice
۴. Ligtvoet
۵. Person fit
۶. Restscore method
۷. Item-splitting method
۸. Proportion matrices
۹. Restscore splitting method

مقیاس بندی مدل‌های موکن برای پاسخ‌های چند ارزشی

مولنار (۱۹۹۷)، تحلیل مقیاس موکن را برای سوال‌های چند ارزشی[۱] گسترش داد. اصول MSA برای سوال‌های چندارزشی مانند سوال‌های دوارزشی است با این تفاوت که برای سوال‌های چندارزشی، مفروضات مدل هم در سطح سوال و هم در سطح گزینه‌های سوال بررسی می‌شود. یک سوال لیکرت با پنج رده پاسخ «کاملاً موافق»، «موافق»، «بی‌نظر»، «مخالف» و «کاملاً مخالف» دارای چهار مرحله[۲] است. در مورد سوال‌های چند ارزشی، هر مرحله دارای یک تابع پاسخ است که به‌عنوان تابع پاسخ مرحله سوال[۳] (ISRF) شناخته می‌شود. ISRF ها احتمال تأیید یک رده‌ی خاص را در صفت نهفته θ نشان می‌دهند.

تعداد ISRF ها همیشه $m-1$ در هر سوال است، m تعداد رده‌های پاسخ در پرسشنامه است؛ بنابراین در یک مقیاس لیکرت ۵ درجه‌ای، هر سوال دارای چهار ISRF است. فرض کنید x_i امتیاز سوال i باشد، $i = ۰,۱, \dots, m$، بنابراین برای یک مقیاس پنج‌درجه‌ای، این به معنای $x_i = ۰,۱,۲,۳,۴$ است. احتمال $x_i=۰$ (کمترین مقدار مقیاس) نادیده گرفته می‌شود زیرا طبق تعریف، این برابر با یک است. مدل همگنی یکنواخت برای سوالات چندارزشی باید برای مراحل سوال رعایت شود؛ یعنی احتمال تأیید رده k یا بالاتر، با افزایش مقادیرθ، کاهشی نداشته باشد. رده‌های پاسخ باید نظم معنی‌داری داشته باشند و سطوح بالاتر، مقدار بیشتری از صفت پنهان را نشان دهند (بقایی، ۲۰۲۱).

فرض یکنواختی برای سوال‌های چندارزشی حاکی از آن است که ISRF ها باید در θ غیرکاهشی باشند. یکنواختی پنهان[۴] یک فرض غیرقابل مشاهده است، یعنی $P(X_i \geq x|\theta)$ در θ غیرکاهشی است. یکنواختی آشکار[۵] یک ویژگی قابل‌مشاهده است به‌عنوان‌مثال $P(X_i \geq x|R = r)$؛ r نمره مانده‌ایست که نماینده θ برای بررسی یکنواختی پنهان قابل‌مشاهده است (شبیه به IIO آشکار که ویژگی قابل‌مشاهده IIO است). برای پاسخ‌های دوارزشی سوال‌های دارای یکنواختی آشکار دلالت بر یکنواختی پنهان دارد اما برای سوال‌های چند ارزشی این‌طور نیست و این رابطه پیچیده‌تر است. شکل ۵ (c) تبعیت از فرض یکنواختی را نشان می‌دهد زیرا ISRF ها با افزایش نمرات مانده، کاهش نمی‌یابند (مولنار، ۱۹۹۷).



شکل ۵. توابع ISRF برای یک سوال ٤ گزینه‌ای

۱. polytomous items
۲. Steps
۳. Item Step Response Function
٤. Latent monotonicity
٥. Manifest monotonicity

مدل یکنواختی مضاعف برای سوال‌های چندارزشی شامل سه فرض: تک‌بعدی بودن، یکنواختی و استقلال موضعی به همراه ISRF های غیرمتقاطع است. به این معنی که احتمال تأیید یک رده معین یا بالاتر روی یک سوال باید دارای نظم نسبی یکسانی در تمام مقادیر θ باشد. ISRFها طبق تعریف در درون یک سوال متقاطع نیستند. مولنار (۱۹۹۷) بیان کرد کهISRF های سوال‌های مختلف نیز باید غیر متقاطع باشند. بااین‌حال برآوردن این شرایط در داده‌های واقعی بسیار دشوار است. علاوه بر این ISRF های ثابت برای سوال‌های مختلف به معنای IRFهای ثابت برای سوال‌ها نیست (سیجتسما، میر و وندرارک ۲۰۱۰).

ازآنجا که ISRF های ثابت برای سوال‌های مختلف به ایجاد امتیازهای ترتیبی مفید کمک نمی‌کنند و در سوال‌های چندارزشی تمرکز بر IRF های کلی ثابت است (سیجتسما، میر و وندرارک ۲۰۱۰). توجه به این نکته مهم است که برای MSA چندارزشی، نسخه ضعیف‌تری از ترتیب تصادفی بر روی متغیر پنهان[۱] (SOL) وجود دارد. برای سوال‌های دوارزشی، اگر مدل MH برقرار باشد از نمره کل X^+ می‌توان برای ترتیب تصادفی افراد روی θ استفاده کرد (سیجتسما و مولنار ،۲۰۰۲). برای سوال‌های چند ارزشی، از دیدگاه نظری، X^+ ممکن است در مرتب کردن افراد روی θ شکست بخورد (همکر، سجیتسما، مولنار و یونکر، ۱۹۹۷)؛ اما وندرآرک (۲۰۰۵) نشان داد که اگر تعداد سوال‌ها بیش از پنج باشد، نقض ترتیب تصادفی به‌ندرت رخ می‌دهد.

وندرآرک و برگسما[۲] (۲۰۱۰) ثابت کردند که برای سوال‌های چند ارزشی شکل ضعیف‌تری از ترتیب تصادفی برقرار است. در سوال‌های چند ارزشی اگر مدل همگنی یکنواخت برقرار باشد، به این معنی است که با استفاده از نمره کل، نمونه را می‌توان به دو گروه تقسیم کرد، به‌طوری‌که افراد دارای بالاترین و کمترین مقدار θ را شناسایی کرد. همکر، سیتسما، مولنار و یونکر (۱۹۹۷) استدلال کردند که در مورد سوال‌های چند ارزشی، نمره کل X^+ پاسخ‌دهندگان را دقیقاً بر روی θ مرتب نمی‌کند. بااین‌وجود، وندرآرک (۲۰۰۰) با استفاده از مطالعات شبیه‌سازی نشان داد که برای اهداف عملی، X^+ به‌درستی افراد را مرتب می‌کند و تحریف‌ها فقط برای پاسخ‌دهندگان با نمره کلی بسیار نزدیک (به‌اندازه یک یا دو امتیاز از هم) رخ می‌دهد.

ده گام برای انجام تجزیه تحلیل مقیاس موکن

مرحله ۱: بررسی داده‌ها. داده‌ها بررسی‌شده و در صورت بروز مشکلات خاص، اقدامات مناسب انجام می‌شود (گام‌های ۱ تا ۳).

گام (۱) کدگذاری مجدد. باید نمرات سوال‌های که به‌صورت منفی[۳] نوشته‌شده‌اند بر اساس ویژگی موردانداره‌گیری مجدداً کدگذاری شوند. به‌طوری‌که برای همه سوال‌ها، نمرات بالاتر به معنای موقعیت بالاتر در مقیاس اندازه‌گیری باشد (سیجتسما و وندرآرک، ۲۰۱۷).

گام (۲) نمرات غیرقابل‌قبول[۴] و داده‌های بدون پاسخ[۵]. با نمرات غیرقابل‌قبول به‌عنوان مقادیر بدون پاسخ (گمشده) برخورد می‌شود. باید درصد کلی سوال‌ها بدون پاسخ را در مجموعه داده‌ها و پاسخ‌دهندگانی که تعدادی سوال را بی‌پاسخ گذاشته یا تعدادی پاسخ نامعقول داده‌اند به دست آورد و از خود پرسید که چرا این حجم از داده‌ها ازدست‌رفته‌اند و چرا برخی پاسخ‌دهندگان خاص تعداد زیادی از سوال‌ها را بی‌پاسخ رها کرده‌اند. اگر به‌عنوان‌مثال بیش از ۱۰٪ از کل داده‌ها ازدست‌رفته‌اند، آیا در طرح مطالعه[۶] یا عبارت[۷] برخی سوال‌ها مشکلی وجود دارد؟ اگر به‌عنوان‌مثال آزمودنی حداقل ۳۰٪ از سوال‌ها را بی‌پاسخ رها کرده است، آیا آن‌ها انجام آزمون را جدی تلقی کرده بودند؟ می‌توان نمرات سوال‌ها بدون پاسخ را با استفاده از روش جاگذاری دوراهه[۸]

---

۱. Stochastic Ordering on the Latent variable
۲. Bergsma
۳. Negatively worded
٤. Inadmissible scores
٥. Missing data
٦. Study design
۷. Wording
۸. Two-way imputation

محاسبه کرد (ونگینکل[۱]، وندرآرک[۲] و سیجتسما، ۲۰۰۷) اما روش‌های دیگری همچون جاگذاری چندگانه (وندرآرک و سیجتسما، ۲۰۰۵) نیز امکان‌پذیر است.

گام (۳) داده‌های پرت[۳]. باید مشخص شود که آیا نمره‌های سوال‌ها واجد شرایط برای داده‌های پرت، دارای الگوی خاصی هستند یا خیر؟ چراکه می‌توان شاهد الگوهای پاسخ غیرمنتظره بود. به‌عنوان نمونه بسیار مشاهده‌شده است که افراد به سوال‌ها دشوار، پاسخ صحیح و به سوال‌ها آسان، پاسخ نادرست می‌دهند. در تجزیه‌وتحلیل داده‌ها توصیه می‌شود که تحلیل یک‌بار بر روی داده‌های کامل و بار دیگر بدون داده‌های پرت انجام گردد. اگر حذف تعداد کمی از داده‌های پرت تا حد زیادی بر نتایج مقیاس‌گذاری تأثیر می‌گذارد، حذف موجه خواهد بود، زیرا نمی‌توان پذیرفت که تنها تعداد اندکی مشاهدات، نتایج مقیاس بندی را تا حد زیادی تحت تأثیر قرار دهند. می‌توان از تعداد خطاهای گاتمن با شاخص $G+$ و ترکیب آن با روش‌های توکی[۴] برای تشخیص داده‌های پرت استفاده کرد (زیجلسترا[۵]، وندرآرک و سیجتسما، ۲۰۱۱).

مرحله ۲: شناسایی مقیاس. باید یک یا چند مقیاس را شناسایی کرد که هر دو مدل همگنی یکنواخت و در صورت نیاز مدل رضایت‌بخش‌تر یکنواختی مضاعف را نیز بررسی کند (گام‌های ۴ تا ۷).

گام (۴) مقیاس‌پذیری. اگر هدف بررسی مقیاس از پیش[۶] تعیین‌شده است، ضرایب H_{jk}، H_j و H^T را برای مقیاس کامل محاسبه می‌شود. اگر هدف بررسی ابعاد[۷] مجموعه داده‌ها است، باید AISP[۸] را c برابر ۰، ۰/۰۵، ۰/۱۰ و ۰/۵۵ اجرا کرد. می‌توان از همکر[۹] و همکاران (۱۹۹۵) برای جستجوی الگوهای مرتبط استفاده کرد. بسته به داده‌ها، الگوی نتایج ممکن باید قبل از رسیدن به بالاترین مقدار c مشخص گردد. AISP به‌صورت تقریبی مقیاسی را مرتب می‌کند که افراد را با استفاده از نمرهٔ کل با خطای نسبتاً کم رتبه‌بندی می‌کند، اما نیاز است تعدادی از سوال‌ها مهم برای تصمیم‌گیری‌های دقیق در نظر قرار گیرند: (مراجعه به گام‌های ۵ تا ۷).

هنگامی‌که تمام مقادیر ضریب مقیاس‌پذیری جفت سوال‌ها مثبت باشد و همه ضرایب مقیاس‌پذیری تک‌تک سوال‌ها بیشتر از ۰/۳۰ باشد، این تعداد سوال، به‌عنوان مقیاس موکن در نظر گرفته می‌شوند (واتسون[۱۰] و همکاران، ۲۰۱۲).

علاوه بر ضرایب فوق، ابعاد آزمون می‌تواند با استفاده از مقادیر حد پایین الگوریتم ژنتیک[۱۱] مورد بررسی قرار گیرد (استرات[۱۲]، وندرآرک و سیجتسما، ۲۰۱۳). حد پایین c حداقل مقدارتمیز(H_j)، برای سوال‌های که مقیاس موکن را تشکیل می‌دهند نشان می‌دهد. تقسیم مجموعه‌ای از سوال‌ها به مقیاس موکن با استفاده از تابع AISP یک روش اکتشافی برای به‌دست آوردن مجموعه‌ای از سوال‌ها است که برخی از انتظارات قابل‌مشاهده اساسی را که با مدل همگنی یکنواخت و قدرت تمیز منطقی همخوانی دارد، برآورده می‌کند (وندرآرک، ۲۰۱۲). به گفته سیجتسما و مولنار (۲۰۰۲)، چنانچه همه سوال‌ها انتخاب‌شده در یک مقیاس (آزمون) دارای ($c\leq$ ۰/۳) باشند، مقیاس یک‌بعدی است. درصورتی‌که مقادیرc افزایش یابد، این سوال‌ها برای قرار گرفتن در مجموعه سوال‌ها مقیاس انتخاب نمی‌شوند.

---

۱. Van Ginkel
۲. Van der Ark
۳. Outliers
۴. Tukey's fences
۵. Zijlstra
۶. Priori
۷. Dimensionality
۸. Automated item selection procedure
۹. Hemker
۱۰. Watson
۱۱. Genetic algorithm
۱۲. Straat

گام (۵) استقلال موضعی. فرض استقلال موضعی بیان می‌کند که اگر پاسخ‌های شخص به سوال‌ها مقیاس، تنها به سطح وی در ویژگی نهفته‌ای که اندازه‌گیری می‌شود، وابسته باشد، پاسخ به یک سوال تحت تأثیر نمرهٔ سایر سوال‌ها قرار نمی‌گیرد. باید تأکید کرد که ردیابی این ویژگی تا حد زیادی حدسی است زیرا استقلال موضعی، پدیده‌ای غیرقابل‌کشف و عملاً غیرقابل تحقق است (پالمگرن[۱] و همکاران، ۲۰۱۸). لرد (۱۹۸۰) می‌گوید وقتی‌که مفروضه تک‌بعدی بودن برقرار است، استقلال موضعی نیز حاصل می‌شود و از این حیث دو مفهوم باهم برابرند (همبلتون، سوامیناتان و راجرز، ۱۹۹۰). بااین‌وجود استرات[۲]، وندرآرک و سیجتسما (۲۰۱۳)، معتقدند که استقلال موضعی را می‌شود با استفاده از روش وابستگی شرطی[۳] بررسی کرد. روش وابستگی شرطی شامل سه شاخص W_1، W_2 وW_3 است که جفت سوال‌های دارای وابستگی (همبستگی) موضعی را علامت‌گذاری می‌کند (سیجتسما و وندرآرک، ۲۰۱۷).

گام (۶) یکنواختی. یکنواختی IRF ها را با استفاده از رگرسیون غیرپارامتری نمره سوال بر روی نمره کل در $J - 1$ سوال دیگر در همان مقیاس بررسی می‌شود (جانکر[۴] و سیجتسما، ۲۰۰۰). تحلیل گرافیکی، این امکان را فراهم می‌آورد که نشان دهد تا چه حد می‌توانیم منحنی مشاهده‌شده را یکنواخت در نظر بگیریم و انحراف موضعی (مکانی) از یکنوایی را می‌توان با استفاده از آزمون آماری مورد آزمون قرار داد (سیجتسما و مولنار، ۲۰۰۲).

گام (۷) رتبه‌بندی نامتغیر سوال‌ها. محققانی که قصد ساخت مقیاسی را دارند که افراد را در یک بعد رتبه‌بندی کند، می‌توانند از این مرحله صرف‌نظر کنند. بااین‌حال اگر نیاز به رتبه‌بندی نامتغیر سوال‌ها وجود داشته باشد، ضریب مقیاس‌پذیری H^T بیانگر درجه‌ای است که پاسخ‌دهندگان به‌صورت نامتغیر به سوال‌ها پاسخ داده‌اند (لیگتوت[۵]، وندرآرک، تی مارولد[۶] و سیجتسما، ۲۰۱۰). فرض IIO از طریق تعداد دفعات نقض مفروضه‌های (عدم تقاطع IRF ها) و روش انتخاب سوال به روش پس‌رو[۷] (سوال‌های را که باید با نقض IIO حذف شوند حذف می‌کند)، موردبررسی قرار می‌گیرد. علاوه بر این شاخص (H^T) فاصله بین IRF ها را تخمین می‌زند. هرچه فاصله بین IRF بیشتر باشد، IIO بیشتر خواهد بود؛ بنابراین آماره H^T به‌عنوان شاخصی از دقت IIO بر اساس معیارهای زیر گزارش می‌شود: ($H^T <$ ۰/۳) به معنای نادرست بودن ترتیب سوال است، (۰/۴ $> H^T \geq$ ۰/۳) به معنی دقت کم، (۰/۵ $> H^T \geq$ ۰/۴) نشانگر دقت متوسط و ($H^T \geq$ ۰/۵) نشانگر دقت بالا (سیجتسما، میر[۸] و وندرآرک، ۲۰۱۰). سرانجام مقدار ملاک[۹] پیشنهادی توسط سیجتسما و مولنار (۲۰۰۲) برای بررسی اندازه اثر یک نقض[۱۰] (تخلف) معنی‌دار با معیارهای زیر بکار می‌رود. اگر (۴۰ $>$ Crit) تخلف جزئی است. اگر (۸۰ $>$ Crit $\geq$ ۴۰)، نقض جدی نیست، اما باید توسط محقق بررسی شود؛ اگر (Crit $\geq$ ۸۰) باشد، نقض جدی است (ون شوور، ۲۰۱۱).

مرحله ۳: خصوصیات (ویژگی‌های) مقیاس. باید ویژگی‌های مقیاس مشخص‌شده در مرحله ۲ را معلوم کرد (مراحل ۸ تا ۱۰).

گام (۸) پایایی[۱۱]. برای ارزیابی میزان پایایی مقیاس، چهار ضریب پایایی قابل برآورد است. اولین ضریب آلفا است. دومین ضریب، آماره لامبدا-۲[۱۲] (سیجتسما، ۲۰۰۹) است که به‌عنوان جایگزین آلفای کرونباخ معرفی‌شده است. سومین ضریب، آمارهٔ سیجتسما- مولنار (MS) است که

---

۱. Palmgren
۲. Straat
۳. Conditional association procedure
۴. Junker
۵. Ligtvoet
۶. Te Marvelde
۷. Backward
۸. Meijer
۹. Criterion
۱۰. Violation
۱۱. Reliability
۱۲. Lambda-۲

به‌عنوان یک برآورد کنندهٔ پایایی با اریبی جزئی (سیجتسما و مولنار، ۲۰۰۲)؛ برای مدل‌های همگنی مضاعف (سیجتسما و مولنار، ۱۹۸۷) کاربرد دارد؛ و چهارمین ضریب، پایایی طبقه پنهان[۱] (LCRC) است که آماره‌ای ناریب از پایایی نمره آزمون (وندرآرک، وندرپالم[۲]، سیجتسما، ۲۰۱۱) محسوب می‌شود.

گام (۹) هنجارها. محققانی که برای اندازه‌گیری افراد به جداول هنجار نیاز دارند، هنجارها و فواصل اطمینان را برای هنجار با استفاده از روش رگرسیون تخمین می‌زنند (اوسترهوز[۳]، وندرآرک و سیجتسما، ۲۰۱۶). در تحقیقات علمی، برای تفسیر عملکرد افراد در آزمون نیازی به هنجاریابی نیست. در این مورد، می‌توان از مرحله ۹ عبور کرد.

گام (۱۰) مقایسه گروه. اگر نمونه شامل زیرگروه‌های معنی‌دار باشد، بررسی ترکیب مقیاس‌ها و ویژگی‌های مقیاس (گام‌های ۴ تا ۹) را می‌توان در بین زیرگروه‌ها نیز تعمیم داد. اگر ترکیب مقیاس و ویژگی‌های مقیاس در گروه‌ها متفاوت است، آگاهی از این تنوع هنگام تفسیر عملکرد آزمون افراد و در پژوهش‌هایی که ویژگی‌های گروه موردمطالعه مانند میانگین نمرات مقیاس و همبستگی نمرات مقیاس با سایر متغیرها موردتوجه است مفید خواهد بود (سیجتسما و وندرآرک، ۲۰۱۷).

روش پژوهش

از آنجا که برای نشان دادن کاربرد MSA در ارزیابی عملکرد آموزشی نیاز به داده‌های دوارزشی است، داده‌های ثانویه پژوهش حاضر، از مطالعه سلطانی نژاد (۱۳۹۶) که هدف آن سنجش میزان آگاهی از تکنیک صحیح اندازه‌گیری فشارخون دانشجویان دو سال آخر دندانپزشکی دانشگاه کرمان (سال تحصیلی ۹۵-۹۶) بر اساس دستورالعمل انجمن قلب آمریکا بوده است، انتخاب شد. جمعیت موردمطالعه در این پژوهش دانشجویان دوره عمومی دو سال آخر دانشکده دندانپزشکی کرمان در سال تحصیلی ۹۵-۹۶ بوده‌اند. روش جمع‌آوری اطلاعات این مطالعه، سرشماری بوده است. بدین نحو که از کلیه دانشجویان که در سال تحصیلی ۹۵-۹۶ دانشجوی سال پنجم یا ششم دندانپزشکی بودند، (مجموعاً ۱۳۸ نفر) خواسته شد تا در موعد مقرر، به سوال‌ها این آزمون، پاسخ دهند. ۴۱/۴ درصد از دانشجویان مذکر و ۵۸/۶ درصد مؤنث بودند. ۵۲/۵ درصد دانشجوی سال پنجم و ۴۷/۵ درصد دانشجوی سال ششم دندانپزشکی بودند. ۲۶/۳ درصد از دانشجویان شهریه‌پرداز و ۷۳/۷ درصد غیرشهریه‌پرداز بودند. ابزار اندازه‌گیری در این پژوهش شامل آزمونی است که ضمن ثبت خصوصیات جمعیت‌شناختی (جنسیت، سال تحصیلی، نحوهٔ ورود به دانشگاه) شامل ۱۴ سوال مربوط به تکنیک صحیح اندازه‌گیری فشارخون بیمار (به‌عنوان نمونه *روی‌هم انداختن پاها هنگام اندازه‌گیری فشارخون ممکن است منجر به کاهش فشارخون شود*) است و بر اساس دستورالعمل انجمن قلب آمریکا طراحی گردیده است. برای هر سوال سه پاسخ (صحیح- غلط و نمی‌دانم) در نظر گرفته‌شده تا دانشجویان پاسخ صحیح را انتخاب کنند. هر جواب صحیح یک نمره مثبت داشته و برای پاسخ غلط و نمی‌دانم امتیازی در نظر گرفته نشده است، بدین ترتیب مجموع نمره آزمون ۱۴ است. بر اساس مطالعه سلطانی نژاد (۱۳۹۶) سنجش روایی محتوایی آزمون با نظرخواهی از ۱۰ نفر از صاحب‌نظران رشته‌های بیماری‌های دهان و جراحی انجام‌شده و روایی آن ۰/۹۹ گزارش‌شده است. برای ارزیابی پایایی آزمون از روش بازآزمایی استفاده گردیده بدین نحو که سوال‌ها آزمون در اختیار ۲۰ نفر از دانشجویان قرارگرفته است و پس از دو هفته مجدداً همان افراد به سوال‌ها پاسخ داده‌اند. ضریب پایایی ۰/۸ محاسبه شده است.

یافته‌ها

برای تحلیل داده‌های آزمون از بسته موکن[۴] در نرم‌افزار R (وندرآرک، ۲۰۱۸) استفاده شده است.

---

۱. Latent class reliability coefficient

۲. Van der Palm

۳. Oosterhuis

۴. Mokken Package

گام اول، دوم و سوم: نتایج بررسی داده‌ها نشان می‌دهد که نمره‌گذاری همه سوال‌ها مثبت و در یک‌جهت است و نیاز به کدگذاری معکوس ندارند. در ضمن دادهٔ بی‌پاسخ و نیز دادهٔ پرت در آن‌ها وجود ندارد و همان‌طور که در جدول ۴ نشان داده‌شده است، مقادیر کجی و کشیدگی در محدودهٔ توزیع نرمال است.

جدول ٤. آماره‌های توصیفی

تعداد سوال‌ها	چولگی	خطای معیار چولگی	کشیدگی	خطای معیار کشیدگی
١٤	-٠/٠٦	٠/٢٤٣	٠/٠٥٥	٠/٤٨١

گام چهارم: برآورد ضرایب مقیاس‌پذیری

ضرایب مقیاس‌پذیری موکن شامل H_{ij}، H_j و H^T با توجه به معیارهای پیشنهادی موکن (۲۰۱۱) موردبررسی قرار گرفتند.

مطابق نتایج جدول ۵، ضرایب مقیاس‌پذیری تک‌تک سوال‌ها و نیز ضرایب مقیاس‌پذیری جفت سوال‌ها (به‌جز در یک مورد) مثبت با خطای معیار کم به‌دست‌آمده است. ضریب مقیاس‌پذیری کل (۰/۴۹۰) نشان می‌دهد که مجموعه ۱۴ سوالی در تشکیل مقیاس موکن در حد متوسط عمل می‌کند. علاوه بر ضرایب فوق، ابعاد آزمون با استفاده از مقادیر حد پایین c=۰/۳ الگوریتم ژنتیک موردبررسی قرارگرفته است. تحلیل داده‌ها (جدول ۵) نشان می‌دهد که به‌جز سوال ۵، سایر سوال‌ها آزمون دارای c<۰/۳ هستند و بنابراین می‌توانند در مجموعه سوال‌ها مقیاس موکن برای سنجش یک صفت تک‌بعدی قرار گیرند. لذا سوال ۵ به‌عنوان کاندیدای حذف در نظر گرفته می‌شود.

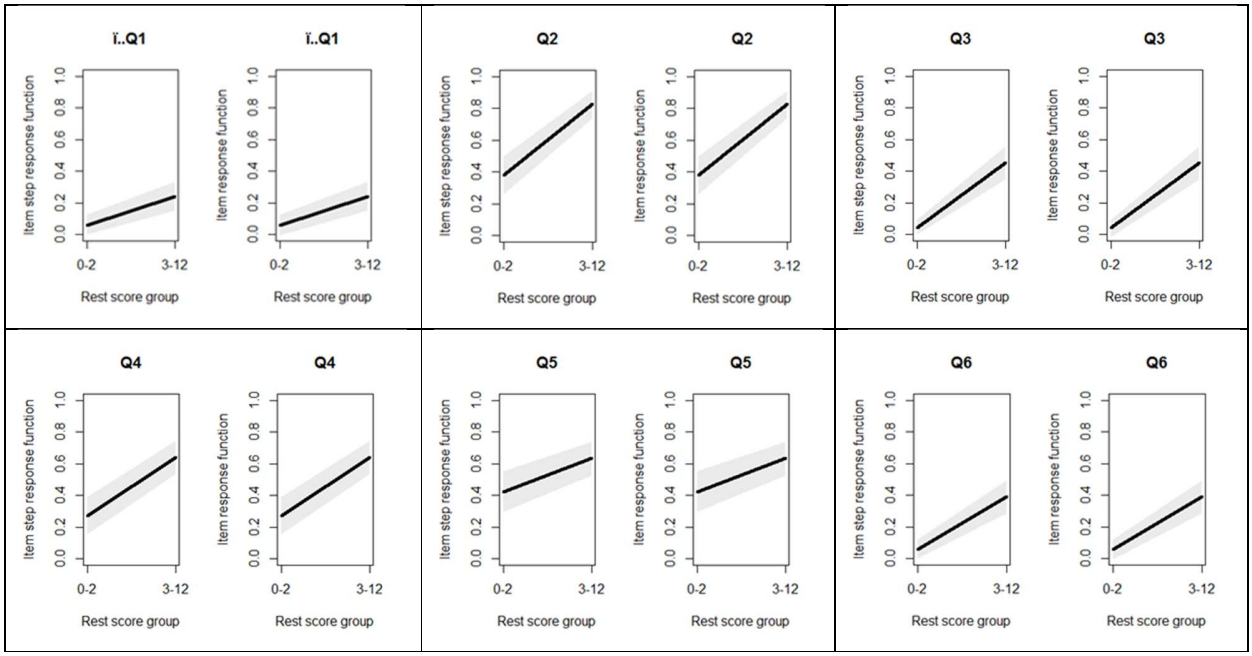
جدول ۵. ضرایب مقیاس‌پذیری تک سوال‌ها

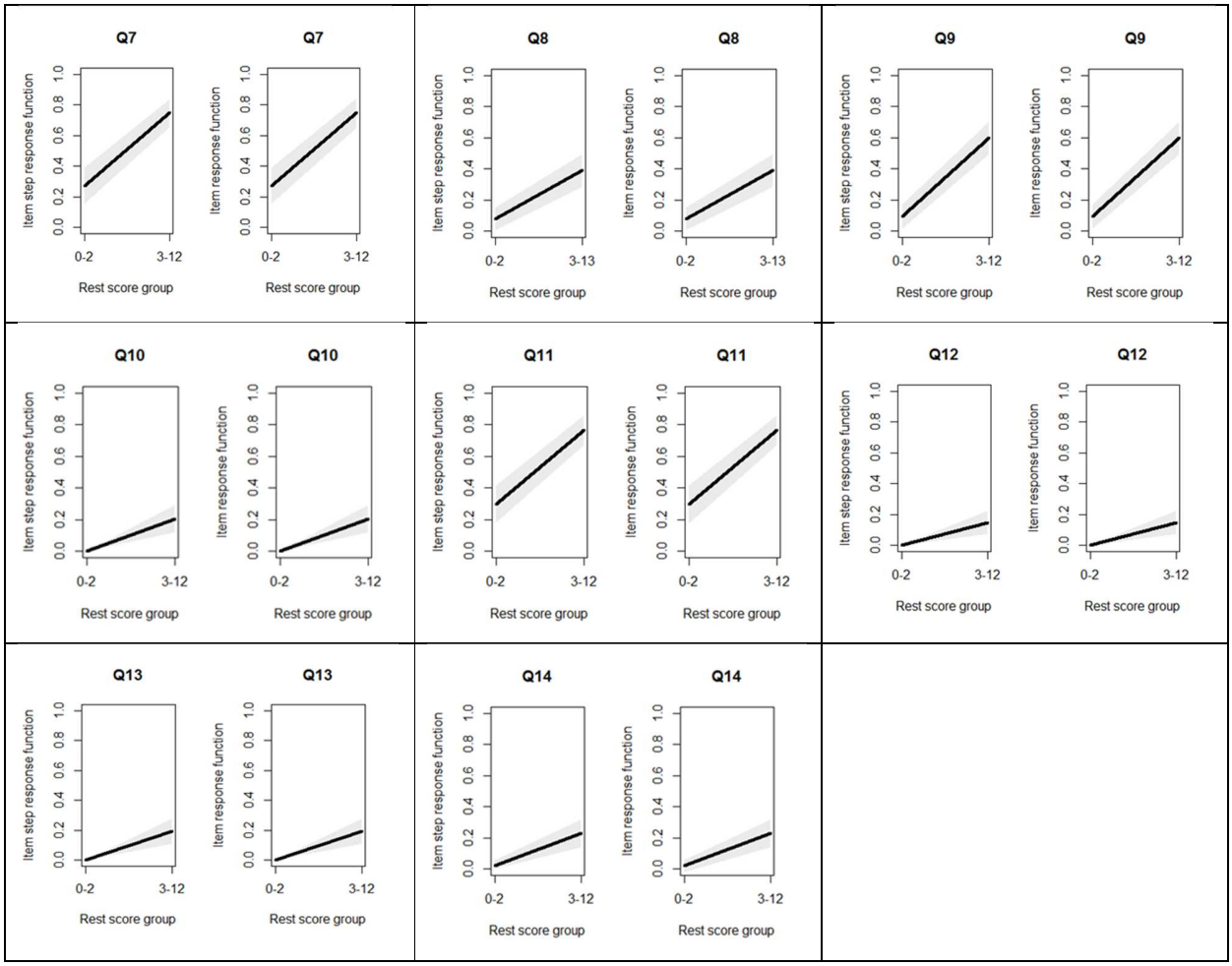
شماره سوال	ضریب مقیاس‌پذیری برآورد شده هر سوال (H_j)	خطای معیار برآورد	ضریب مقیاس‌پذیری کل ($H_.$)	الگوریتم ژنتیک تعیین تعداد زیر مقیاس با c= ۰/۳
١	٠/٤٣٢	٠/٠٨٣	٠/٤٩٠	١
٢	٠/٥١٥	٠/٠٥٧		١
٣	٠/٦٢٠	٠/٠٥٤		١
٤	٠/٤٧٢	٠/٠٥٨		١
٥	٠/٢٣٤	٠/٠٧٥		.
٦	٠/٥٠٢	٠/٠٦٥		١
٧	٠/٤٨٨	٠/٠٥٤		١
٨	٠/٤٩٢	٠/٠٦٨		١
٩	٠/٦١٣	٠/٠٤٧		١
١٠	٠/٤٥٧	٠/٠٦٣		١
١١	٠/٥٠٢	٠/٠٥٢		١

شماره سوال	ضریب مقیاس‌پذیری برآورد شده هر سوال (H_j)	خطای معیار برآورد	ضریب مقیاس‌پذیری کل ($H_.$)	الگوریتم ژنتیک تعیین تعداد زیر مقیاس با c= ۰/۳
۱۲	۰/۵۸۷	۰/۰۵۷		۱
۱۳	۰/٤۳۲	۰/۰۸۱		۱
۱٤	۰/۵۳۲	۰/۰۷٤		۱

گام پنجم: استقلال موضعی. ازآنجاکه مفروضهٔ تک‌بعدی بودن برقرار است، استقلال موضعی نیز حاصل‌شده است، چون استقلال موضعی به‌طور خودکار از تک‌بعدی بودن پیروی می‌کند و مفروضهٔ جداگانه‌ای به‌حساب نمی‌آید. البته برخی از متخصصان نظیر همبلتون و سوامیناتان معتقدند استقلال موضعی به معنای تک‌بعدی بودن نیست (لرد، ۱۹۸۱؛ ترجمه دلاور و یونسی، ۱۳۹۱). بااین‌وجود با توجه به اینکه ردیابی استقلال موضعی تا حد زیادی حدسی و عملاً غیرقابل تحقق است (پالمگرن و همکاران، ۲۰۱۸) و وابستگی موضعی در موقعیت‌های سنجش پیشرفت تحصیلی و استعداد ایجاد مشکل نمی‌کند (امبریتسون و رایس، ۲۰۱۳) در اکثر تحلیل‌ها این مفروضه به‌طور فنی چک نمی‌شود.

گام ششم: یکنواختی. بررسی گرافیکی فرض یکنواختی توابع پاسخ سوال‌ها (شکل ۶) تأیید می‌کند که در همه سوال‌های آزمون، فرض یکنواختی (غیر کاهشی بودن) برقرار است. لذا نتایج تحلیل تا به اینجا نشان می‌دهد که مدل همگن یکنواخت موکن با داده‌ها برازش دارد. در شکل ۳ نمودارهای غیر کاهشی توابع پاسخ سوال‌های آزمون آورده شده است.



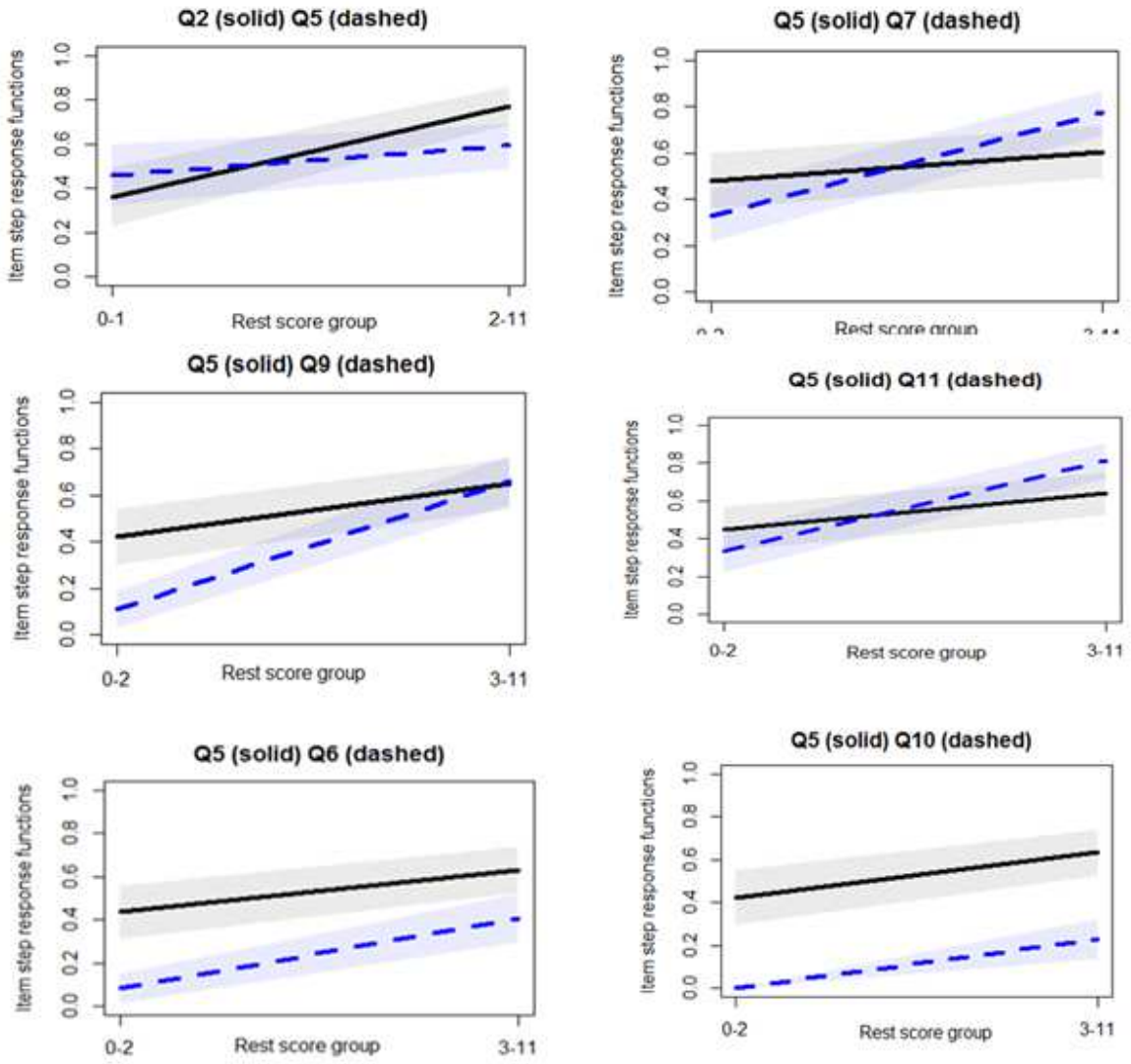


شکل ٦. نمودارهای غیرکاهشی توابع پاسخ سوال‌های آزمون

گام هفتم: رتبه‌بندی نامتغیر سوال‌ها: فرض IIO را با در نظر گرفتن تعداد دفعات نقض مفروضه‌های (عدم تقاطع IRF ها) و روش انتخاب سوال به روش پس‌رو موردبررسی قرار گرفت. برای بررسی اندازه اثر یک نقض (تخلف) معنی‌داری آن با شاخص مقدار ملاک بررسی گردید. نتایج جدول ۶، نشان می‌دهد که به‌عنوان مثال تابع پاسخ سوال ۲، با ۱۳ سوال دیگر مقایسه شده است که از این تعداد، تنها در یک مورد IRF سوال ۲ با سوال دیگری تلاقی دارد که آن مقدار تلاقی معنی‌دار نیست. برای سوال ۷ یک مورد نقض معنی‌دار وجود دارد، اما چون شاخص ملاک آن زیر ۸۰ به‌دست‌آمده است، لذا سوال ۷ هم در مرحله انتخاب پس‌رو، کاندیدای حذف معرفی نشده است؛ اما برای سوال ۵، چهار مورد نقض مفروضه وجود دارد (شکل ۷) که از این تعداد، یک مورد معنی‌دار است. مقدار ملاک هم تنها برای سوال ۵ بیشتر از ۸۰ به‌دست‌آمده است که نشان می‌دهد نقض جدی مفروضه در سوال ۵ روی داده است و لذا نتایج انتخاب سوال به روش پس‌رو نشان می‌دهد که برای داشتن مجموعه سوال‌هایی با مدل همگنی مضاعف، سوال ۵ باید حذف شود. ضریب H^T (۰/۴۵) است که نشان‌دهندهٔ یک ترتیب قابل‌قبول است. هرچند که نتایج تکرار تحلیل با حذف سوال ۵، نشان داد که ضریب H^T برای اطلاعات باقیمانده همچنان ۰/۴۵ است، اما حذف سوال ۵، شاخص مقیاس‌پذیری کل را به ۰/۵۴(در حد قوی) افزایش داد. همچنین ضرایب پایایی برای مجموعه ۱۳سوالی از ۰/۸۶ تا ۰/۹۱ متغیر بود (جدول شماره ۷). در کل نتایج نشان می‌دهند که شواهد برازش مدل یکنواختی مضاعف نیز وجود دارد. لذا از مجموعه سوال‌ها (با حذف سوال ۵) می‌توان هم برای رتبه‌بندی افراد و هم برای رتبه‌بندی سوال‌ها استفاده کرد.

جدول ٦. شاخص‌های بررسی مفروضه تغییرناپذیری ترتیب سوال‌ها

شماره سوال	تعداد مقایسه‌های جفتی برای هر سوال	تعداد موارد نقض مفروضه	تعداد نقیضه‌های معنی‌دار	crit	انتخاب پس رو		H^T
					مرحله ۱	مرحله ۲	
۲	۱۳	۱	۰	۲۳	۰	۰	۰/٤٤٩٩
۱۱	۱۳	۱	۰	۲۹	۰	۰	
۷	۱۳	۱	۱	٥٦	۰	۰	
٥	۱۳	٤	۱	۱۱٥	۱	NA	
٤	۱۳	۱	۰	۱۲	۰	۰	
۹	۱۳	۰	۰	۰	۰	۰	
۳	۱۳	۱	۰	٥	۰	۰	
۸	۱۳	۱	۰	۱۲	۰	۰	
٦	۱۳	۰	۰	۰	۰	۰	
۱	۱۳	۰	۰	۰	۰	۰	
۱٤	۱۳	۰	۰	۰	۰	۰	
۱۰	۱۳	۰	۰	۰	۰	۰	
۱۳	۱۳	۰	۰	۰	۰	۰	
۱۲	۱۳	۰	۰	۰	۰	۰	



شکل ۷. نمونه‌ای از نمودارهای جفتی تابع پاسخ سوال ۵ با سوال‌های دیگر

گام هشتم: پایایی

برای ارزیابی میزان پایایی مقیاس، چهار ضریب پایایی برآورد کردیم. جدول شماره ۷، مقادیر ضرایب پایایی را هم برای کل سوال‌ها (۱۴ سوال) و هم بعد از حذف سوال ۵، برای مجموعه ۱۳ سوالی نشان می‌دهد. مقایسه ضرایب پایایی نشان می‌دهد که همه شاخص‌های پایایی پس از حذف سوال ۵، افزایش‌یافته‌اند.

جدول ۷. مقادیر ضرایب پایایی

پایایی مجموعه ۱۴ سوالی				پایایی مجموعه ۱۳ سوالی (با حذف سوال ۵)			
MS	λ_2	alpha	LCRC	MS	λ_2	alpha	LCRC
۰/۸۵۶	۰/۸۶۱	۰/۸۵۳	۰/۹۰۶	۰/۸۰۶	۰/۸۶۷	۰/۸۶۱	۰/۹۱۲

شکل ۶ نشان می‌دهد که به‌عنوان‌مثال تابع پاسخ سوال ۵ با سوال‌های ۲، ۷، ۹ و ۱۱ تلاقی دارد که از این تعداد، تنها در یک مورد مقدار تلاقی معنی‌دار است و برای مثال IRF سوال ۵ با سوال‌های ۶ و ۱۰ تلاقی ندارد. ازآنجاکه در پژوهش حاضر برای تفسیر عملکرد افراد در آزمون نیاز به هنجاریابی نداریم و مقایسه گروه‌ها جزو اهداف نبوده است، گزارش گام‌های نهم و دهم ضروری نیست.

بحث و نتیجه‌گیری

امروزه آزمون‌ها، مقیاس‌های درجه‌بندی، پرسشنامه‌ها و تمامی ابزارهای اندازه‌گیری جزئی جداناپذیر از زندگی بشر شده‌اند. این بحث تا حدی حائز اهمیت است که افراد از بدو تولد با استفاده از ابزارهای مختلف مورد سنجش قرار می‌گیرند و نتایج حاصل از این سنجش‌ها در تصمیم‌گیری‌های متعددی مورداستفاده قرار می‌گیرد. با احتساب این نقش برای آزمون‌ها در زندگی فردی و اجتماعی افراد جامعه، درک درست از نمره‌های حاصل از این ابزار در طول تاریخ همیشه موردتوجه بوده است و نظریه‌های مختلف ازجمله نظریه کلاسیک و نظریه سوال–پاسخ هر یک به‌نوبه خود درصدد بهبود کیفیت اندازه‌گیری‌ها بوده‌اند (تیلور، ۲۰۱۳ ترجمه یونسی، ۱۳۹۸). هرچند که با روی کار آمدن برنامه‌های متعدد و کامپیوترهای پیشرفته، گرایش متخصصان اندازه‌گیری، در تحلیل سوال و ساخت ابزار به سمت مدل‌های سوال –پاسخ بیشتر شده است، اما پیچیدگی‌های محاسباتی زیربنایی آن هنوز به قوت خود باقی است. نیرومندی و اهمیت نظریه سوال– پاسخ و اینکه IRT ازجمله ابزار قدرتمند در ساخت و تجزیه تحلیل آزمون‌ها و مقیاس‌ها به شمار می‌آید، بر هیچ‌کس پوشیده نیست. شاید با توجه به پیشرفت در حوزه‌های مختلف و روی کار آمدن کامپیوترهای پیچیده بتوان اذعان کرد که عصر نظریه کلاسیک به پایان رسیده است. امروزه بیشتر تحقیقات موجود با استفاده از روش‌های روان‌سنجی مدرن مبتنی بر نظریه سوال–پاسخ پارامتری است. IRT پارامتری شامل تعدادی از روش‌های اندازه‌گیری قدرتمند است، اما برای استفاده از این روش‌ها چندین معیار باید برآورده شوند و لذا استفاده از آن‌ها بسیار پیچیده است؛ اما روش‌های نظریه سوال –پاسخ غیرپارامتری (NIRT) محدودیت کمتری نسبت به مدل‌های IRT پارامتری دارند و بعضی از روش‌های NIRT ممکن است ساده‌تر و بصری‌تر از روش‌های پارامتری IRT باشند؛ مانند نظریه کلاسیک آزمون، معیار اصلی در NIRT نمره جمع است، اما روش‌های NIRT در روش انتخاب سوال‌ها، نحوهٔ ارزیابی کیفیت مقیاس و غیره از روش‌های کلاسیک متفاوت است (سیجتسما و مولنار، ۲۰۰۲). مدل‌های اندازه‌گیری مانند CTT یا تحلیل عاملی فرض می‌کنند که همه سوال‌ها دارای وزن یکسان هستند (توزیع فراوانی یکسانی دارند). هر زمان که این فرض نقض شود مانند سوال‌های که مقیاس تجمعی را تشکیل می‌دهند، نتیجه غیرواقعی که می‌تواند بروز کند، این است که سوال‌ها به اندازه کافی همگن نیستند تا یک متغیر پنهان واحد را اندازه بگیرند. در این شرایط، یک مزیت عمده مدل‌های IRT، ازجمله تحلیل مقیاس موکن نسبت به مدل‌های CTT این است که مدل‌های IRT در معرفی پارامترهای مدل برای سوال‌ها، صریحاً در نظر می‌گیرند که سوال‌ها دارای وزن یا (اهمیت[1]) متفاوت هستند. دومین مزیت تجزیه‌وتحلیل مقیاس موکن نسبت به مدل‌هایCTT در تأکید دقیق بر تناسب مدل نهفته است. تمام ضرایب H_{ij} (و بنابراین تمام همبستگی‌های زوجی) باید مثبت باشند و هر سوال باید به اندازه کافی با بقیه همگن باشد. این الزامات منجر به ابزار اندازه‌گیری می‌شود که با استانداردهای بالاتر قابلیت اطمینان (پایایی) و همگنی مطابقت دارند تا ابزارهایی که فقط در یک تحلیل پایایی استاندارد بررسی‌شده‌اند. سومین و کاربردی‌ترین مزیت روش مقیاس‌گذاری موکن نسبت به سایر روش‌های اندازه‌گیری، روش

[1]. Popularity

جستجوی خوشه‌ای سلسله مراتبی[۱] «از پایین به بالا» است که برای یافتن زیرمجموعه‌ای با حداکثری از سوال‌های همگن بکار می‌رود. این روش به‌ویژه در تحقیقات اکتشافی با هدف ایجاد ابزار اندازه‌گیری جدید به محقق کمک می‌کند که حتی در صورت وجود تعداد محدود سوال‌ها، کاندیداهای جدید متغیرهای پنهان را کشف کند. سرانجام تجزیه‌وتحلیل مقیاس موکن یک مدل IRT است که می‌تواند با موفقیت برای تعداد اندکی از سوال‌ها استفاده شود. مولنار (۱۹۹۷) نشان داد که وقتی تعداد سوال‌ها نسبتاً کم باشد، نتایج تحلیل مقیاس موکن و تحلیل مقیاس راش، اغلب به نتایج یکسانی منجر می‌شوند (ون شوور، ۲۰۰۳).

References

Abdelhafez, A. M. . (۲۰۰۷). Postgraduate Research Students' Knowledge and Attitudes towards Good Supervisory Practice at the University of Exeter. *Online Submission.*

Abdelhamid, G. S., Gómez-Benito, J., Abdeltawwab, A. T., Abu Bakr, M. H., & Kazem, A. M. (۲۰۲۰). A Demonstration of Mokken Scale Analysis Methods Applied to Cognitive Test Validation Using the Egyptian WAIS-IV. *Journal of Psychoeducational Assessment, 38*(٤), ٤٩٣-٥٠٦.

Andrich, D. (۱۹۸۲). An index of person separation in latent trait theory, the traditional KR. ۲۰ index, and the Guttman scale response pattern. *Education Research and Perspectives, 9*(۱), ۹٥-۱۰٤.

Baghaei, P. (۲۰۲۱). *Mokken Scale Analysis in Language Assessment.* Germany: Münster.

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (۲۰۰۱). . (۲۰۰۱). Fitting item response theory models to two personality inventories: Issues and insights. Multivariate Behavioral Research, ۳٦(٤), ٥۲۳-٥٦۲.

Cronbach, L. J. (۱۹٥۱). Coefficient alpha and the internal structure of tests. *psychometrika, 16*(۳), ۲۹۷-۳۳٤.

Emons, W. H., Sijtsma, K., & Pedersen, S. S. (۲۰۱۲). Dimensionality of the Hospital Anxiety and Depression Scale (HADS) in cardiac patients: comparison of Mokken scale analysis and factor analysis. *Assessment, 19(3), 337-353., 19*(۳), ۳۳۷-۳٥۳.

Engelhard Jr, G. (۲۰۰۸). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement, 6*(۳), ۱٥٥-۱۸۹.

Guttman, L. (۱۹٤۷). On Festinger's evaluation of scale analysis. . *Psychological Bulletin, 44*(٥), ٤٥۱.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (۱۹۹۱). *Fundamentals of item response theory* (Vol. ۲). Sage.

Hemker, B. T., K. Sijtsma, I. W. Molenaar, and B. W. Junker. (۱۹۹۷). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, ۳۳۱-۳٤۷.

Hemker, B. T., Sijtsma, K., & Molenaar, I. W. (۱۹۹٥). Selection of unidimensional scales from a multidimensional item bank in the polytomous Mokken I RT model. *Applied Psychological Measurement, 19*(٤), ۳۳۷-۳٥۲.

Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (۱۹۹۷). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62*(۳), ۳۳۱-۳٤۷.

Junker, B. W., & Sijtsma, K. (۲۰۰۰). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24*(۱), ٦٥-۸۱.

Ligtvoet, R. V. (۲۰۱۱). Polytomous latent scales for the investigation of the ordering of items. *Psychometrika*(۲), ۲۰۰-۲۱٦.

Ligtvoet, R., Van der Ark, L. A., Te Marvelde, J. M., & Sijtsma, K. (۲۰۱۰). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement, ۷۰*(٤), ٥۷۸-٥۹٥.

[۱]. Hierarchical clustering search

Loevinger, J. (۱۹٤۸). The technic of homogeneous tests compared with some aspects of" scale analysis" and factor analysis. *Psychological bulletin, 45*(٦), ٥۰۷.

Lord, F. M. (۱۹۸۰). Applications of item response theory to practical testing problems.(A. Delavar ،J. Younesi, Trans.). Roshd Press.

Lumsden, J. (۱۹٥۷). A factorial approach to unidimensionality. *Australian Journal of Psychology, 9*(۲), ۱۰٥-۱۱۱.

Meijer, R. R., & Baneke, J. J. (۲۰۰٤). Analyzing psychopathology items: a case for nonparametric item response theory modeling. *Psychological methods, 9*(۳), ۳٥٤.

Meijer, R. R., Tendeiro, J. N., & Wanders, R. B. (۲۰۱٤). *The use of nonparametric item response theory to explore data quality.* Routledge.

Mokken, R. J. (۱۹۷۱). A theory and procedure of scale analysis. The Hague, The Netherlands: Mouton. *MokkenA Theory and Procedure of Scale Analysis1971, 62*(۳), ۳۳۱-۳٤۷.

Mokken, R. J. (۱۹۷۱). *A theory and procedure of scale analysis: With applications in political research* (Vol. (Vol. ۱)). Walter de Gruyter.

Molenaar, I. W. (۱۹۹۷). *Nonparametric models for polytomous responses. In Handbook of modern item response theory.* New York, NY: Springer.

Mooij, T. (۲۰۱۲). A Mokken scale to assess secondary pupils' experience of violence in terms of severity. *Journal of Psychoeducational Assessment,, 30*(٥), ٤۹٦-٥۰۸.

Oosterhuis, H. E., van der Ark, L. A., & Sijtsma, K. (۲۰۱۷). Standard errors and confidence intervals of norm statistics for educational and psychological Tests. *psychometrika, 82*(۳), ٥٥۹-٥۸۸.

Palmgren, P. J., Brodin, U., Nilsson, G. H., Watson, R., & Stenfors, T. (۲۰۱۸). Investigating psychometric properties and dimensional structure of an educational environment measure (DREEM) using Mokken scale analysis–a pragmatic approach. *BMC medical education, 18*(۱), ۱-۱٦.

Reise, S. P., & Waller, N. G. (۲۰۰۹). Item response theory and clinical measurement. *Annual review of clinical psychology, 5*, ۲۷-٤۸.

Sijtsma, K. (۲۰۰۹). Correcting fallacies in validity, reliability, and classification. *International Journal of Testing, 9*(۳), ۱٦۷-۱۹٤.

Sijtsma, K., & Meijer, R. R. (۲۰۰۷). ۲۲ Nonparametric Item Response Theory and Special Topics. *Handbook of statistics, 26*, ۷۱۹-۷٤٦.

Sijtsma, K., & Molenaar, I. W. (۱۹۸۷). Reliability of test scores in nonparametric item response theory. *Psychometrika, 52*(۱), ۷۹-۹۷.

Sijtsma, K., & Molenaar, I. W. (۲۰۰۲). *Introduction to nonparametric item response theory* (Vol. Vol. ٥). sage.

Sijtsma, K., & van der Ark, L. A. (۲۰۱۷). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology, 70*(۱), ۱۳۷-۱٥۸.م

Sijtsma, K., and I. W. Molenaar. (۲۰۰۲). Introduction to Nonparametric Item Response. *Thousand Oaks, CA: Sage*.

Sijtsma, K., Meijer, R. R., & van der Ark, L. A. (۲۰۱۱). Mokken scale analysis as time goes by: An update for scaling practitioners. *Personality and Individual Differences, 50*(۱), ۳۱-۳۷.

Soltaninejad, Fatemeh. (۱۳۹٦). Assessing the knowledge of Kerman dental students about the correct way to measure the patient's blood pressure according to the standard of the American Heart Association in the ۹٥-۹٦ academic year. (Unpublished doctoral dissertation). Kerman University of Medical Sciences and Health Services, School of Dentistry.

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (۲۰۱۳). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification, 30*(۱), ۷۵-۹۹.

Taylor, C. S. (۲۰۱۳). Validity and validation.(J. Younesi, Trans.). Allameh Tabatabai University Press.

Ünlü, A. (۲۰۰۷). Nonparametric item response theory axioms and properties under nonlinearity and their exemplification with knowledge space theory. *Journal of Mathematical Psychology, 51*(٦), ۳۸۳-٤۰۰.

Van der Ark, L. A. (۲۰۱۲). New developments in Mokken scale analysis in R. *Journal of Statistical Software, 48*(٥), ۱-۲۷.

Van der Ark, L. A. (۲۰۰٥). Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika, 70*(۲), ۲۸۳-۳۰٤.

Van der Ark, L. A. (۲۰۱۰). A note on stochastic ordering of the latent trait using the sum of polytomous item scores. *Psychometrika, 75*(۲), ۲۷۲-۲۷۹.

Van Der Ark, L. A., & Sijtsma, K. (۲۰۰٥). The effect of missing data imputation on Mokken scale analysis. *New developments in categorical data analysis for the social and behavioral sciences*, ۱٤۷-۱٦٦.

Van der Ark, L. A., van der Ark, M. L. A., & Suggests, M. A. S. S. (۲۰۱۸). Package 'mokken'.

Van der Ark, L. A., van der Palm, D. W., & Sijtsma, K. (۲۰۱۱). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement, 35*(٥), , ۳۸۰-۳۹۲.

Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (۲۰۰۷). Multiple imputation of item scores in test and questionnaire data, and influence on psychometric results. *Multivariate Behavioral Research, ٤۲*(۲), ۳۸۷-٤۱٤.

Van Schuur, W. H. (۲۰۰۳). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis, 11*(۲), ۱۳۹-۱٦۳.

Van Schuur, W. H. (۲۰۱۱). *Ordinal item response theory: Mokken scale analysis*. Thousand Oaks: SAGE.

Watson, R. v. (۲۰۱۲). Item response theory: how Mokken scaling can be used in clinical practice. *Journal of clinical nursing, 21*(۱۹pt۲۰), ۲۷۳٦-۲۷٤٦.

Wind, Stefanie A. (۲۰۱۷). An Instructional Module on Mokken Scale Analysis. *Educational Measurement*, ۱–۱۷.

Zijlstra, W. P., van der Ark, L. A., & Sijtsma, K. (۲۰۱۱). Outliers in questionnaire data: Can they be detected and should they be removed? *Journal of Educational and Behavioral Statistics, 36*(۲), ۱۸٦-۲۱۲.