Educational Measurement and Evaluation Studies

National Organization for Educational Testing

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

An Investigation of the Evaluators' Ratings of the Performance Exams in the Field of Arts Using Multi-Faceted Rasch Model

Sara Hajatpour Qala Rudkhani [1], Balal Izanloo[2]

1. Master of Educational Research, Kharazmi University, Tehran, Iran, Email : hajatpoursara94@gmail.com
2.Assistant Professor, faculty of Psychology and Education, Kharazmi University, tehran, iran; (Corresponding Author), Email: izan.b@khu.ac.ir

Article Info

Article Type:
Research Article

Received: 2023.02.04
Received in revised form:2023.05.08
Accepted:2023.06.02
Published online:
2023.06.26

ABSTRACT

Objective: The present study was done in order to examine the severity/leniency and the central tendency level of raters in scoring of performance tests performed by National Organization for Educational Testing (NOET).

Methods: For this purpose, the secondary data in Sketch Architecture Test (1396 and 1397 solar, respectively with 5437 and 7459 people), Industrial design test (1396 solar, 1365 people), Music recognition test (1397 solar, 569 people), playwriting test (1396 solar, 97 people). The data were analyzed by classical methods and many-faceted Rasch models and the results extracted.

Results: The results from classical methods show that in both Sketch Architecture Tests, raters' consistency is generally acceptable, but in other tests (Industrial design, music cognition and playwriting) homogeneity is low. Raters' consensus is low in all five examined tests. Results from many-facet Rasch models show that in both Sketch Architecture Tests rater's severity and use of lower scores of rating scale effects are present, but as expected, there was not any effect for central tendency. Unfortunately, due to the nature of incorrect data collection designs in Industrial design, music cognition and playwriting tests analysis with many-facet Rasch models was not possible.

Conclusion: Based on findings it is recommended that when global performance tests are evaluated by the NOET organization raters, firstly; the proper design for evaluating have been selected and used, and secondly; to prevent the effect of severity or leniency and agreement (consensus) reduction between raters, the training of them for scoring performance tests to be considered

Keywords: Multi-faceted Rush models, severity, leniency, central tendency, performance tests tests.

Cite this article: Hajatpour Qala Rudkhani, Sara; Izanloo, Balal (2023). An Investigation of the Evaluators' Ratings of the Performance Exams in the Field of Arts Using Multi-Faceted Rasch Model. *Educational Measurement and Evaluation Studies*. 13(42):88-112 pages. DOI: 10.22034/EMES.2023.528161.2244

Publisher: National Organization of Educational Testing (NOET)

مطالعات اندازه‌گیری و ارزشیابی آموزشی
سازمان سنجش آموزش کشور
شاپا چاپی: ۲۸۶۵-۲۴۷۶ شاپا الکترونیکی: ۰۹۴۲-۲۷۸۳

بررسی رتبه‌دهی ارزیابان آزمون‌های عملکردی سراسری (طراحی صنعتی، شناخت موسیقی، نمایش عروسکی، طراحی معماری و اسکیس معماری) بر اساس روش‌های کلاسیک و مدل‌های چندوجهی راش

سارا حاجت پورقلعه رودخانی[۱]، بلال ایزانلو[۲]

۱. کارشناس ارشد تحقیقات آموزشی، دانشگاه خوارزمی، تهران، ایران. رایانامه: hajatpour.sara98@gmail.com
۲. استادیار دانشکده روانشناسی و علوم تربیتی، دانشگاه خوارزمی، تهران، ایران؛ (نویسنده مسئول)، رایانامه: izan.b@khu.ac.ir

اطلاعات مقاله

نوع مقاله:
مقاله پژوهشی

دریافت: ۱۴۰۱/۱۱/۱۵
اصلاح: ۱۴۰۲/۰۲/۱۸
پذیرش: ۱۴۰۲/۰۳/۱۲
انتشار : ۱۴۰۲/۰۴/۰۵

چکیده

هدف: پژوهش حاضر به منظور بررسی میزان جدیت/تساهل و گرایش به مرکز ارزیابان در نمره‌گذاری آزمون‌های عملکردی سراسری سازمان سنجش انجام شد.

روش پژوهش: برای این منظور از داده‌های ثانویه آزمون‌های طراحی معماری سال‌های ۱۳۹۶ (۵۴۳۷ نفر)، اسکیس معماری ۱۳۹۷ (۷۴۵۹ نفر)، طراحی صنعتی سراسری سال ۱۳۹۶ (۱۳۶۵نفر)، موسیقی سال ۱۳۹۷ (۵۶۹ نفر) و نمایش عروسکی سال ۱۳۹۷ (۹۷ نفر) استفاده شد. داده‌ها با استفاده از روش‌های کلاسیک و مدل‌های چند وجهی راش تحلیل و نتایج استخراج شد.

یافته‌ها: در هر دو آزمون طراحی معماری، در کل میزان همسانی (مطابقت نسبی بین درجه‌بندی ارزیابان) قابل قبول، ولی در آزمون‌های طراحی صنعتی، شناخت موسیقی و نمایشنامه‌نویسی میزان همسانی پایین است. میزان اجماع (توافق) در هر پنج آزمون مورد بررسی نیز پایین است.

نتیجه‌گیری: نتایج حاصل از مدل‌های چندوجهی نشان داد اثر سخت‌گیری ارزیاب و استفاده از نمره‌های کرانه‌ای پایین در نمره‌گذاری، در هر دو آزمون طراحی معماری وجود داشت، ولی مطابق انتظار اثر گرایش به مرکز وجود نداشت. به دلیل استفاده از طرح‌های جمع‌آوری داده‌های نامناسب (مثلا در همه پاسخ‌ها یا تکالیف یک آزمون، هر داور فقط ۲ مورد مجزا از هم را ارزیابی می‌کند و هیچ همپوشی بین موارد ارزیابی شده توسط ارزیابان وجود ندارد تا با هم به صورت درست مقایسه شوند یا این که در یک آزمون هر تکلیف یا سوال توسط داوران متفاوت ارزیابی می‌شود) در آزمون‌های طراحی صنعتی، شناخت موسیقی و نمایشنامه‌نویسی امکان تحلیل با مدل‌های چندوجهی وجود نداشت. با توجه به یافته‌ها توصیه می‌شود به هنگام ارزیابی آزمون‌های عملکردی سراسری، اولا از طرح مناسب برای ارزیابی استفاده شود و دوما با آموزش ارزیابان در زمینه نمره‌گذاری آزمون‌های عملکردی از تاثیر عواملی مثل جدیت یا تساهل و کاهش توافق جلوگیری به عمل آید.

واژه‌های کلیدی: مدل‌های چند وجهی راش، جدیت، تساهل، گرایش مرکزی، آزمون‌های عملکردی سراسری.

استناد: حاجت پور قلعه رودخانی، سارا؛ ایزانلو، بلال (۱۴۰۲). بررسی رتبه‌دهی ارزیابان آزمون‌های عملکردی سراسری (طراحی صنعتی، شناخت موسیقی، نمایش عروسکی، طراحی معماری و اسکیس معماری) بر اساس روش‌های کلاسیک و مدل‌های چندوجهی راش. مطالعات اندازه‌گیری و ارزشیابی آموزشی، ۱۳ (۴۲): ۸۸-۱۱۲ صفحه. DOI: 10.22034/EMES.2023.528161.2244

ناشر: سازمان سنجش آموزش کشور

مقدمه

تاثیر ویژگی‌های رفتاری ارزیاب بر رتبه‌بندی عملکرد افراد در آزمون‌های مختلف از نظر تجربی روشن است (انگلهارد[۱]، ۱۹۹۴). نمره‌گذاری انجام شده توسط ارزیابان باید از عینیت لازم برخوردار بوده و تاثیر سوگیرهای مختلف مثل تساهل/جدیت[۲]، گرایش مرکزی[۳]، اثر هاله‌ای[۴] و محدودیت دامنه[۵] تا حد امکان کاهش یابد. اصطلاح تساهل معمولا برای تمایل ارزیاب به درجه‌بندی بالاتر و اصطلاح جدیت به تمایل ارزیاب به درجه‌بندی پایین‌تر از نقطه میانی مقیاس اندازه‌گیری، اشاره دارد. نتیجه آنی تساهل یا جدیت آن است که رتبه‌ها در سطح بالا یا پایین هر مقیاس اندازه‌گیری جمع می‌شوند. در گرایش مرکزی ارزیابان از طبقه وسط مقیاس اندازه‌گیری بیش از حد استفاده می‌کنند، درحالی که از کاربرد طبقات کرانه‌ای اجتناب می‌نمایند. منظور از اثر هاله‌ای آن است که برداشت کلی فرد ارزیاب از فرد و شایستگی‌های او بر ارزیابی از صفات مختلف وی تاثیر دارد. محدودیت دامنه برای موقعیت‌هایی استفاده می‌شود که در آن درجه‌بندی‌ها اطراف هر نقطه‌ای در پیوستار درجه‌بندی جمع می‌شوند. فرقی نمی‌کند که این نقطه، نقطه بالایی مقیاس درجه‌بندی باشد یا نقطه پایین و یا نقطه وسطی. دسته‌ای از اثرات ارزیاب که معمولا کمتر مورد اشاره قرار می‌گیرند عبارتند از: ۱-عدم دقت[۶] ۲- خطای منطقی[۷] ۳-خطای مقابل[۸] ۴- تاثیرات رفتارها، عقاید، نگرش‌ها و ویژگی‌های شخصیتی ارزیاب ۵- تاثیرات مشخصات مربوط به ارزیاب یا ارزیابی‌شونده ۶- خطای مجاورت[۹] ۷- خطای تاخر یا تقدم[۱۰] ۸- اثرهای ترتیب[۱۱] (مایفُرد و ولف[۱۲]، ۲۰۰۳).

در بافت نظریه کلاسیک از روش‌های مختلفی برای بررسی عملکرد ارزیابان استفاده می‌شود. شاخص اجماع پایایی بین ارزیابان[۱۳]، که به توافق بین ارزیابان[۱۴] نیز معروف است، به این که ارزیابان مستقل تا چه اندازه درجه‌بندی یکسانی از یک آزمون‌شونده یا شیء خاص ارائه می‌دهند اشاره دارد (مطابقت مطلق درجه‌بندی‌ها). در مقابل شاخص همسانی پایایی بین ارزیابان به این نکته اشاره دارد که ارزیابان مستقل تا چه اندازه به آزمون‌شونده‌ها یا اشیایی که قرار است رتبه‌بندی شوند رتبه نسبی یکسانی اختصاص می‌دهند (مطابقت نسبی درجه‌بندی‌ها) (استملر و سای[۱۵]، ۲۰۰۸؛ تینسلی و وایس[۱۶]، ۲۰۰۰). برای مثال ممکن است یک ارزیاب به آزمون شوندگان امتیازهایی بدهد که به طور مداوم یک یا دو نمره کمتر از امتیازی است که یک رتبه‌دهنده دیگر به همان آزمون شوندگان اعطا می‌کند. در این حالت ترتیب نسبی آزمون شوندگان (مطابقت نسبی) برای هر دو ارزیاب یکسان و برآورد همسانی آنها بالا خواهد بود، ولی ارزیاب‌ها در هیچ موردی به توافق دقیقی (اجماع) نرسیده‌اند (ایکس، ۲۰۱۱). شاخص‌های اجماع که معمولا در عمل استفاده می‌شوند عبارتند از: شاخص توافق دقیق[۱۷] و کاپای وزن‌دار کوهن[۱۸]. توافق دقیق برابر است با تعداد

---

1. Engelhard
2. Leniency, Severity
3. Central Tendency
4. Halo Effect
5. Restriction Of Range
6. Inaccuracy
7. Logical Error
8. Contrast Error
9. Proximity Error
10. Recency Or Primacy Error
11. Order Effects
12. Myford & Wolfe
13. consensus index of interrater reliability
14. Interrater Agreement
15. Stemler & Tsai
16. Tinsley & Weiss
17. Exact Agreement Index
18. Cohens Weigted Kappa

مواردی که رتبه‌های یکسانی دریافت کرده‌اند تقسیم بر تعداد کل مواردی که توسط هر دو ارزیاب درجه‌بندی شده‌اند. شاخص کاپا توافق بین ارزیابان (دو ارزیاب) را به خاطر توافقی که صرفا به دلیل حدس انتظار می‌رود اصلاح می‌کند. شاخص‌های همسانی عبارتند از: ضریب همبستگی گشتاوری[1] و تاو b کندال[2] . ضریب همبستگی گشتاوری پیرسون که به آن r پیرسون نیز می‌گویند میزان رابطه خطی بین رتبه‌بندی دو داور را نشان می‌دهد. تاو b کندال میزان مطابقت بین دو مجموعه رتبه‌ای که عملکرد آزمون شونده‌ها را رتبه‌بندی کرده نشان داده و رتبه‌های گره‌دار را لحاظ می‌کند. هر دو شاخص پایایی در دامنه صفر و یک قرار دارند، مقادیر بالای این دو شاخص همبستگی یا مطابقت قوی‌تر بین درجه‌بندی‌ها را نشان می‌دهد. در شاخص توافق دقیق مقادیر ۰/۷ بالا محسوب می‌شود و مقدار ۰/۱ بسیار پایین در نظر گرفته می‌شود. مقادیر بین ۰/۴ تا ۰/۵ در سنجش‌های پرمخاطره جزو مقادیر بسیار پایین محسوب شده و رضایت بخش نیستند. مقادیر کاپای کوهن وزن‌دار قابل قبول، نزدیک ۰/۷ به بالا قرار دارند. اگرچه در اکثر موارد نتیجه شاخص‌های اجماع و همسانی یکسان است، ولی در موارد استثنایی ممکن است مقادیر همسانی بالا ولی مقادیر اجماع پایین‌تر از حد قابل قبول باشند (فلایس و همکاران[3] ،۲۰۰۳). معمولا وقتی براساس همبستگی پیرسون در مورد همسانی بین ارزیابان قضاوت کنیم مقادیر همسانی قابل قبول است ولی وقتی براساس تائوی کندال در مورد همسانی قضاوت ارزیابان قضاوت شود میزان همسانی چندان قابل قبول نیست، چرا که ضریب تائوی کندال وجود گره در رتبه‌بندی‌ها را لحاظ می‌کند به همین دلیل مقدار همسانی کاهش می‌یابد.

یکی از روش‌های کلاسیک بررسی انواع سوگیری و نقش آنها در رتبه‌بندی و نمره‌گذاری استفاده از نظریه تعمیم‌پذیری[4] است، که جزو رویکردهای کلاسیک اندازه‌گیری محسوب شده و به کمک مبانی تحلیل واریانس، سهم هر عامل موثر در نمره‌گذاری مشخص می‌شود (اسمیت و کولِکویچ[5]، ۲۰۰۴). امروزه علاوه بر نظریه تعمیم‌پذیری از رویکرد مدل‌های چندوجهی راش نیز برای تحلیل رتبه‌های اختصاص یافته به تکالیف و فعالیت‌های افراد استفاده می‌شود. ملاک انتخاب بین این دو رویکرد تحلیلی روشن است. اگر هدف برآورد میزان شباهت نمره‌های خام مشاهده شده گروهی از دانش آموزان و نمره‌های خامی که گروه مشابهی ممکن است تحت شرایط ایده‌آل به دست آورند باشد، نظریه تعمیم‌پذیری مفید خواهد بود. ولی در صورتی که برای هر فرد برآورد نمره‌ای که تا حد امکان از ویژگی‌های عوامل ایجادکننده آن مستقل باشد اهمیت دارد، باید از رویکرد مدل‌های چندوجهی راش[6] (MFRM) استفاده کرد (کیم و ویلسون[7]، ۲۰۰۹).

عموما سنجش‌های عملکردی نه تنها وجوه آزمون‌شونده و سوال (یا فعالیت) را در بر می‌گیرد، بلکه وجوه دیگری مثل، ارزیاب‌ها، معیار نمره‌گذاری، مصاحبه کننده‌ها و احتمالا موارد بیشتری را شامل می‌شود. پیشنهادات اولیه برای بسط مدل اصلی راش به منظور در نظر گرفتن همزمان سه یا چند وجه (عامل‌های آزمایشی) توسط میکو[8] (۱۹۶۹، ۱۹۷۰) و کمف[9] (۱۹۷۲) انجام شد. این رویکرد که مبتنی بر مدل‌های خصیصه پنهان است و در بافت مدل‌های راش شکل گرفته و گسترش پیدا کرده، به طبقه‌ای از مدل‌های اندازه‌گیری اشاره دارد که برای تحلیل همزمان چند متغیر که به طور بالقوه بر نتایج سنجش تاثیر دارند مناسب است. از زمان اولین بیان نظری گسترده مدل‌های چندوجهی توسط لیناکر[10] (۱۹۸۹) تعداد کاربردهای اساسی اندازه‌گیری چندوجهی راش در حوزه‌های سنجش زبان، اندازه‌گیری روانشناسی و تربیتی، علوم بهداشت و بسیاری حوزه‌های دیگر به سرعت در حال افزایش

1. Product-Moment Correlation
2. Kendalls Tau-b
3. Fleiss
4. generalizability theory
5. Smith & Kulikowich
6. multi-facet Rasch models(MFRM)
7. Kim and Wilson
8. Micko
9. Kempf
10. Linacre

است. نقل قول زیر از کتاب لیناکر (۱۹۸۹) نیروهای پشت پرده موثر که باعث رشد سریع رویکرد MFRM در تحلیل و ارزیابی سنجش‌های مبتنی بر ارزیاب شده‌اند را برجسته می‌کند. «یک طراح آزمون وظیفه‌شناس متوجه می‌شود که روش متداول تصمیم-گیری برای عملکردهای درجه‌بندی شده به‌طور مستقیم بر اساس نمره‌های خام برای آزمودنی‌هایی‌که با ارزیاب‌های سخت‌گیر مواجه‌اند ناعادلانه است. همچنین ممکن است زمانی که آزمون‌شونده‌های فاقد صلاحیت در یک رشته حیاتی صرفا به این خاطر که با ارزیابان سهل‌گیر مواجه شده‌اند گواهی (مدرک) دریافت کرده‌اند. این نوع شرایط واقعی این سوال را مطرح می‌کند که چگونه می‌توان اندازه‌های منصفانه و معناداری از درجه‌بندی‌های رتبه‌ای به ناچار مشکوک تهیه کرد؟

برای درک مدل‌های چندوجهی راش گریز کوتاهی به مدل اصلی راش و مدل‌های چند مقوله‌ای[۱] راش لازم است. فرم لگاریتم شانس مدل اصلی راش با گرفتن لگاریتم طبیعی (لگاریتم در مبانی عدد نپر که تا رقم اعشار برابر ۲/۷۱۸ است) از نسبت شانس (بخش داخل پرانتز در رابطه زیر که نسبت احتمال پاسخ درست به سوال i توسط فرد n به احتمال پاسخ نادرست به همان سوال توسط همان فرد را نشان می‌دهد) به دست می‌آید. این مدل برای سوال‌هایی که پاسخ به آنها در نهایت به صورت نادرست (صفر) و درست (یک) طبقه‌بندی می‌شود مناسب است.

$$\mathrm{In}\left[\frac{P(x_{ni}=1)}{P(x_{ni=0})}\right] = \theta_n - \beta_i \qquad (۱)$$

لگاریتم طبیعی نسبت شانس را لُجیت[۲] می‌گویند، که خلاصه عبارت"واحد لگاریتم شانس[۳] " است. تحت این مدل، لجیت یک تابع خطی ساده از پارامتر توانایی فرد n (θ_n) و پارامتر دشواری سوال i (β_i) است. هم توانایی آزمون شونده و هم دشواری سوال، در این مقیاس بیان شده است. به‌طور کلی، لُجیت واحد اندازه‌گیری مقیاس مربوط به هر پارامتر مشخص شده در مدل راش است. با توجه به مدل نشان داده شده در معادله ۱، یک لُجیت برابر فاصله‌ای در طول مقیاس اندازه‌گیری است که شانس موفقیت را تقریبا ۲/۷۱۸ افزایش می‌دهد، که همان مقدار عدد e است (لیناکر و رایت[۴]، ۱۹۸۹؛ لودلو و هیلی[۵]، ۱۹۹۵). مدل دوارزشی راش که پایه اصلی سایر مدل‌های راش است را می‌توان به بیش از دو طبقه مرتب شده گسترش داد. سوال‌هایی که طبقات پاسخ آنها ذاتا دارای بیش از دو طبقه مرتب شده است را سوال‌های چندمقوله‌ای گفته و آزمون یا مقیاسی که حاوی این‌گونه سوال‌ها باشد را مقیاس‌های درجه‌بندی می‌گویند، به‌طوری که با اجرای آنها داده‌های چند مقوله‌ای حاصل می شود. مثلا سوال‌های نوع لیکرت[۶] با طبقات پاسخ مرتب شده ۴ تایی (و گاهی ۵ تایی) که نگرش افراد در خصوص موضوعات مختلف را اندازه‌گیری می‌کنند دارای طبقات کاملا مخالف[۷]، مخالف[۸]، موافق[۹] یا کاملا موافق[۱۰] هستند. در بافت رویکرد راش به اندازه‌گیری، برای پرداختن به داده‌های حاصل از اجرای مقیاس‌های درجه‌بندی، مدل‌های چندمقوله‌ای راش گسترش یافته‌اند. روشن است که این مدل‌ها به مدل دو مقوله‌ای اصلی راش پارامترهایی اضافه می‌کنند. هدف از این‌کار توصیف کارکرد مقیاس‌های درجه‌بندی حاوی سوال‌های چند مقوله‌ای است (آندریچ[۱۱] ، ۲۰۰۵a؛ امبرتسون و رایز[۱۲]، ۲۰۰۰؛ استینی و نرینگ[۱۳]، ۲۰۰۶؛ پنفیلد[۱۴]، ۲۰۱۴). یکی از مدل‌های چند مقوله‌ای راش، مدل مقیاس درجه‌بندی[۱۵]

---

1. Polytomous
2. logit
3. Log Odds Unit
4. Linacre & Wright
5. Ludlow & Haley
6. Category Likert
7. Strongly Disagree
8. Disagree
9. Agree
10. Strongly Agree
11. Andrich
12. Embretson & Reise
13. Ostini & Nerin
14. Penfield
15. Rating Scale Model

(RSM) است. مدل RSM برای بازنمایی دشواری نسبی انتقال از یک طبقه مقیاس درجه‌بندی به طبقه پاسخ بعدی پارامتر آستانه[1] را اضافه می‌کند. به بیان دقیق‌تر پارامتر آستانه یا ضریب طبقه[2]، τ_k، جایی در بعد پنهان است که احتمال مشاهده طبقات پاسخ مجاور k و k-1 برابر است. به عبارت دیگر، به شرط آن که آزمودنی در یکی از این دو طبقه پاسخ مجاور قرار داشته باشد، τ_k یک نقطه انتقال روی بعد پنهان است که در آن احتمال آنکه آزمودنی به یکی از دو طبقه مجاور پاسخ دهد برابر ۵۰٪ است. این نقاط انتقال را آستانه‌های راش–آندریچ[3] می‌گویند (بوند و فوکس[4]، ۲۰۱۵؛ لیناکر، ۲۰۰۶a، ۲۰۱۴b؛ آندریچ، ۱۹۹۸، ۲۰۰۵a). فرم لگاریتم شانس مدل RSM برابر است با:

$$\ln\left[\frac{P_{nik}}{P_{nik-1}}\right] = \theta_n - \beta_i - \tau_k \qquad (۲)$$

در اینجا P_{nik} احتمال آن که آزمون‌شونده n به طبقه k سوال i پاسخ دهد است؛ P_{nik-} احتمال آن که آزمون‌شونده n به طبقه k-1 سوال i پاسخ دهد؛ k طبقه پاسخ یک مقیاس درجه‌بندی با m+1 طبقه است (یعنی، k=0,…,m)؛ τ_k دشواری پاسخ به طبقه k (نسبت به طبقه k–۱) است. دشواری سوال چندمقوله ای i، β_i، جایی بر روی بعد پنهان تعریف می‌شود که پایین‌ترین و بالاترین طبقات دارای احتمال یکسان هستند. مدل RSM فرض می‌کند همه سوال‌های (مقیاس‌های درجه‌بندی) موجود در آزمون دارای یک مجموعه پارامتر آستانه برابر هستند. بنابراین، مدل فقط زمانی باید استفاده شود که سوال‌ها ساختار مقیاس درجه‌بندی مشترکی دارند. یعنی، زمانی که سوال‌ها تعداد طبقات پاسخ یکسانی دارند و دشواری نسبی بین طبقات در بین سوال‌ها یکسان است. زمانی که سوال‌ها از نظر تعداد طبقات پاسخ متفاوتند یا زمانی که انتظار می‌رود دشواری نسبی بین طبقات پاسخ از یک سوال به سوال دیگر تغییر می‌کند، مدل امتیاز جزئی[5] (PCM) جایگزین مناسبی است. مدل PCM برای هر سوال پارامترهای آستانه جداگانه‌ای برآورد می‌کند، که اجازه می‌دهد ساختار مقیاس درجه‌بندی هر سوال منحصر به‌فرد باشد. فرم لگاریتم شانس مدل امتیاز جزئی برابر است با:

$$\ln\left[\frac{P_{nik}}{P_{nik-1}}\right] = \theta_n - \beta_i - \tau_{ik} \qquad (۳)$$

در اینجا P_{nik} احتمال آن که آزمون شونده n به طبقه k سوال i پاسخ دهد و P_{nik-1} احتمال آن که آزمون شونده n به طبقه k-1 سوال i پاسخ دهد. k طبقه پاسخ مقیاس درجه‌بندی است که $m_i + 1$ طبقه دارد (یعنی، $k=0,\ldots,m_i$). τ_{ik} دشواری پاسخ به طبقه k در سوال i نسبت به طبقه k-1 است (مسترز[6]، ۱۹۸۲، ۲۰۱۰).

در یک موقعیت سنجش شامل ارزیابان، با استفاده از یک مقیاس درجه‌بندی (مثل، یک مقیاس چهار طبقه‌ای) برای ارزیابی کلی کیفیت عملکرد آزمون‌شونده حداقل با دو وجه قابل تمایز ارزیاب‌ها و آزمون‌شونده‌ها وجود دارند. اگر آزمون‌شونده‌ها به هر یک از فعالیت‌ها پاسخ دهند، و ارزیابان درجه عملکرد آزمون‌شونده را در هر فعالیت به‌طور جداگانه درجه‌بندی کنند، پس لازم است وجه فعالیت نیز در نظر گرفته شود. یعنی آزمون‌شونده‌ها، فعالیت‌ها و ارزیاب‌ها یک موقعیت سه وجهی را تشکیل می‌دهند. مدل اندازه-گیری چندوجهی راش مناسب این موقعیت در زیر ارایه شده است.

$$\ln\left[\frac{P_{nljk}}{P_{nljk-}}\right] = \theta_n - \delta_I - \alpha_j - \tau_k \qquad (۴)$$

---

1. Threshold
2. Category Coefficient
3. Rasch Andrich Thresholds
4. Bond & Fox
5. Partial Credit Model
6. Masters

در این معادله P_{nljk} = احتمال اینکه آزمون‌شونده n در فعالیت l از ارزیاب j رتبه k دریافت کند، P_{nljk-1} = احتمال اینکه آزمون‌شونده n در فعالیت l از ارزیاب j رتبه k-1 دریافت کند، θ_n = توانایی آزمون‌شونده n، δ_l = سطح دشواری فعالیت l، α_j = سختگیری ارزیاب j و τ_k = دشواری دریافت رتبه k نسبت به رتبه k-1 است. همان‌طور که در معادله ۴ نشان داده شده مدل MFRM اساسا یک مدل جمع‌پذیر خطی است که بر تبدیل لجستیک رتبه‌های مشاهده شده به مقیاس لجیت استوار است ("جمع‌پذیر" دراینجا به معنی" ترکیب شده به وسیله قوانین جمع و تفریق" است). با استفاده از واژگان آماری استاندارد، می‌توان تبدیل لجستیک نسبت احتمال‌های طبقات متوالی (یعنی، لگاریتم شانس) را به عنوان متغیر وابسته با وجوه مختلف، مثل آزمون‌شونده‌ها، فعالیت‌ها، ارزیاب‌ها که به لحاظ مفهومی به عنوان متغیرهای مستقل که بر لگاریتم‌های شانس تاثیرگذار هستند در نظر گرفت (هیز[1]، ۱۹۹۴؛ مایرز، ول و لورچ[2]، ۲۰۱۰). پارامتر آستانه، τ_k، چگونگی برخورد با داده‌های رتبه‌ای را نشان می‌دهد. این پارامتر مشخص می‌کند که برای تمام عناصر هر وجه مدل مقیاس رتبه‌ای[3] (RSM) باید استفاده شود.

برای مثال، در یک تحلیل، با مقیاس چهار طبقه‌ای طوری برخورد می شود که همه ارزیاب‌ها آن را درک کرده‌اند و هر طبقه مقیاس رتبه‌ای را به شیوه بسیار مشابه استفاده کرده‌اند. درخصوص وجه فعالیت، این موضوع به این معنی است که یک درجه‌بندی خاص، مثل"۲" در فعالیت ۱، معادل با درجه‌بندی"۲" در فعالیت ۲ و در سایر فعالیت‌های موجود در سنجش فرض می‌شود. به بیان دقیق‌تر، پارامترهای آستانه به طور هم‌زمان در بین ارزیاب‌ها، فعالیت‌ها و آزمون شونده‌های درجه‌بندی برآورد می شوند؛ یعنی، مجموعه‌ای از پارامترهای آستانه که ساختار مقیاس درجه‌بندی را تعریف می‌کنند برای همه ارزیاب‌ها (و نیز برای همه فعالیت‌ها و آزمون‌شونده‌ها) یکسان است.

به جای این کار، پارامتر آستانه را می‌توان طوری مشخص کرد تا ساختارهای مقیاس درجه‌بندی در بین عناصر یک وجه خاص متغیر باشد، در این صورت مدل امتیاز جزئی[4] (PCM) سه وجهی حاصل می‌شود. به تغییر در معنی عبارات استفاده شده در مدل RSM سه‌وجهی در مقایسه با عبارت‌های استفاده شده در مدل RSM اصلی توجه کنید. مدل RSM سه وجهی پاسخ آزمودنی به یک طبقه خاص مقیاس درجه‌بندی را مدل سازی نمی‌کند، بلکه این مدل نوع RSM پاسخ ارزیاب به یک طبقه خاص مقیاس ارزیابی را بر اساس آنچه که وی معتقد است عملکرد آزمون‌شونده را مشخص می‌سازد مدل‌سازی می‌نماید. به زبان دیگر در معادله ۴، پارامتر آستانه، τ_k، به دشواری پاسخ دادن به طبقه k ام (نسبت به طبقه k-1) مقیاس درجه‌بندی اشاره ندارد، بلکه به دشواری دریافت پاسخ در طبقه k (نسبت به k-1) اشاره دارد. بنابراین مهم است که بین دشواری که پاسخ به یک طبقه خاص نشان می‌دهد و دشواری مشاهده شده در آن طبقه تمایز قائل شویم (لیناکر و رایت، ۲۰۰۲).

به‌طور کلی، وقتی داده‌هایی که باید مدل سازی شوند حاوی درجه‌بندی‌ها باشند، مشخص کردن RSM چندوجهی (یا PCM چند وجهی) به این پیش فرض که ارزیابان از طبقات مختلف مقیاس درجه‌بندی استفاده کرده‌اند اشاره دارد. پس می‌توان بحث اصلی را به‌صورت زیر بیان کرد: آیا این منطقی است که فرض کنیم که همه ارزیاب‌ها از مقیاس به روش یکسان استفاده می‌کنند؟ آیا ارزیاب‌ها بر روی تفاوت‌های عملکردی به‌دست آمده در طبقات مجاور توافق دارند، یا آیا ممکن است برخی ارزیابان کم و بیش از سبک خاص خود برای انتصاب معنی به طبقات پیروی کنند؟ با افزایش تجارب حرفه‌ای، هر ارزیاب معمولا شیوه یا فرایند خاصی برای رتبه‌بندی فعالیت اتخاذ می‌کند، شیوه‌ای که برای انطباق با کار روزانه ارزیاب تکامل یافته، اما ممکن است از شیوه سایر ارزیابان تا حد زیادی متفاوت باشد. آشنایی ارزیاب‌ها با مقیاسی که استفاده می شود مطمئنا به افزایش موافقت ارزیاب در استفاده از مقیاس کمک می‌کند، اما اینکه در هر نمونه خاص چه میزان توافق می‌تواند به‌دست آید در واقع یک سوال بی‌پاسخ باقی می‌ماند.

---

1. Hays
2. Myers, Well & Lorch
3. Rating Scale Model
5. Partial Credit Model

از دیدگاه اندازه‌گیری، شیوه اصولی برای پاسخ‌گویی به این سوال مستلزم اجرای تحلیل PCM چندوجهی و مقایسه نتایج درجه‌بندی طبقه در بین ارزیابان است. همان‌گونه که در معادله ۴ نشان داده شده است، میزان سخت‌گیری ارزیاب (α) به وضوح مدل‌سازی می‌شود. بنابراین تحلیل بر اساس مدل، برآورد میزان سخت‌گیری هر ارزیاب را نشان می‌دهد. به‌طور کلی، میزان سخت‌گیری ارزیاب زمانی وجود دارد که ارزیاب‌ها درجه‌بندی‌هایی را فراهم کنند که به طور ثابت نسبت به سایر ارزیابان یا درجه‌بندیهای ملاکی تعیین شده (یعنی اجماع رتبه‌بندی‌ها که تو سط گروهی از ارزیابان خبره به سطوح خاصی از توانایی آزمودنی فراهم شده) سخت‌تر باشند. به طور ویژه در معادله ۴، و نیز سایر معادلات پیچیده‌تر مدل‌های چندوجهی، پارامتر α_j میزان سخت‌گیری ارزیاب j، را مدل‌سازی می‌کند. یعنی، هر چه مقدار این پارامتر بزرگ‌تر باشد، درجه‌بندی پیش‌بینی شده کمتر خواهد شد. به بیان دیگر، این مدل نشان می‌دهد که ارزیابان سخت‌گیر یا جدی، تمایل دارند به آزمون شوندهها نمره‌های پایین‌تری اختصاص دهند، که در نتیجه توانایی آزمون‌شونده کمتر می‌شود. برعکس ارزیابان آسان‌گیر تمایل دارند نمره‌های بالاتری به آزمودنی‌ها بدهند. بنابراین، اگر کسی علاقه به مدل‌سازی سهل‌گیری ارزیاب داشته باشد، پارامتر α_j باید (به جای اینکه از معادله مدل کم شود) به معادله مدل اضافه شود.

به خاطر هزینه و به دلایل بهره‌وری، به ویژه در بافت سنجش‌های مقیاس بزرگ، معمولا همه ارزیاب‌ها عملکردهای همه آزمون‌شوندهها را ارزیابی نمی‌کنند. غالبا، طرح سنجش به شیوه درجه‌بندی دوگانه[1] محدود می‌شود؛ یعنی فقط دو ارزیاب از میان گروه بزرگی از ارزیاب‌های نسبتا شایسته به طور مستقل مجموعه یکسانی از عملکردهای آزمون شوندهها را ارزیابی می‌کنند. در چنین موقعیت‌هایی، ماتریس داده‌هایی که همه نمرات داده شده به آزمون شوندهها را در بر دارد ناقص است؛ یعنی تعداد زیادی داده گمشده وجود خواهد داشت. به شرط اینکه طرح درجه‌بندی دارای پیوندهای کافی بین عناصر وجوه موجود باشد، توانایی آزمون‌شونده، میزان سخت‌گیری ارزیاب و پارامترهای دشواری تکلیف[2]، یا هر پارامتری که در یک مدل خاص مشخص شده را می‌توان هنوز براساس چهارچوب رایج مرجع برآورد کرد. به عبارت دیگر، مدل‌های راش در برابر داده‌های گمشده مقاوم هستند، چرا که این مدل‌ها فقط برای نقاط داده‌های مشاهده شده ارزیابی می‌شوند. به همین دلیل در مدل‌های راش نیازی به محاسبه یا تعدیل داده‌های مشاهده نشده نیست. زمانی که بیشتر از سه وجه وجود داشته باشد، ساخت طرح درجه‌بندی مناسب می‌تواند کاملا یک فعالیت چالش برانگیز باشد. برای مطالعه بیشتر در این خصوص به اِکس (۲۰۱۱) مراجعه شود.

آزمون‌های عملکردی سراسری رشته‌های مختلف مقاطع گوناگون، مثل گروه هنر و معماری که هر ساله توسط سازمان سنجش استفاده می‌شوند از اهمیت زیادی برخوردارند، چرا که آینده تحصیلی و در نهایت شغلی شرکت کنندگان در این آزمون‌ها تا حد زیادی به نتایج این آزمون‌ها وابسته است. متاسفانه تا به امروز کاری پژوهشی جدی در خصوص این آزمون‌ها صورت نگرفته است. چون ارزیابان این آزمون‌ها را درجه‌بندی و نمره‌گذاری می‌کنند، لازم است کیفیت درجه‌بندی‌های حاصل از ارزیابی بررسی، مشخص و کیفیت نمره‌گذاری ارزیابان این آزمون‌ها با روش‌های آماری مناسب و مختلف بررسی شود. با توجه به مباحث مطرح شده، هدف پژوهش حاضر بررسی نمره‌گذاری ارزیابان آزمون‌های عملکردی رشته هنر بر اساس رویکردهای کلاسیک و نیز مدل چندوجهی راش است. از این رو اهداف این پژوهش عبارت است از: ۱) بررسی میزان اجماع (توافق) و همسانی ارزیابان ۲) بررسی میزان جدیت/تساهل ارزیابان در رتبه‌بندی آزمون رشته هنر در مقطع کارشناسی ارشد ۳) بررسی میزان گرایش به مرکز ارزیابان در فرایند رتبه‌دهی.

مبانی نظری و پیشینه پژوهش

استفاده از مدل‌های چندوجهی برای بازخورد انفرادی در تصحیح بعدی مصححان تاثیر دارد. به این صورت که وقتی مصححان با استفاده از مقیاس رتبه‌ای، بخش‌های نوشتاری و گفتاری آزمون شغلی انگلیسی را (که در استرالیا برای پذیرش پزشکان مهاجر به این کشور برگزار می‌گردد) مورد ارزیابی قرار دادند، بازخورد انفرادی با استفاده از نتایج مدل‌های چندوجهی راش، باعث کاهش

[1]. Double-Rating
[2]. task difficulty parameters

چهار خطای سخت‌گیری، سهل‌گیری، منظم بودن، و ارزیابی کلی قبل و بعد از بازخورد انفرادی شده است. اگرچه مصححان دید مثبتی نسبت به بازخورد انفرادی دارند (ناک[1]، ۲۰۱۱).

آموزش ارزیابان، منجر به مقیاس‌های دارای ثبات بیشتر بین مصححین می‌شود و سوگیری در استفاده از طبقه‌بندی‌های مقیاس رتبه‌بندی را کاهش می‌دهد (وولف و مکوی[2]، ۲۰۱۲). به علاوه، از آنجا که ریشه کن کردن کامل تنوع مصححین تقریباً غیرممکن است، حتی اگر آموزش کاربردی باشد، بهتر است روش آموزش به عنوان روندی در نظر گرفته شود که مصححین را بیشتر سازگار و متکی به خود می‌سازند (پایایی درونی مصحح‌ها) به جای آنکه به یکدیگر متکی باشند (پایایی بین مصحح‌ها). آموزش باعث می‌شود کیفیت رتبه‌بندی مصححین با تجربه و بی‌تجربه پس از آموزش بهبود یابد؛ با این حال، مصححین بی‌تجربه ثبات بالاتر و سوگیری کمتری دارند. از این‌رو، هیچ گونه شواهدی وجود ندارد که مصححین بی‌تجربه تنها به دلیل عدم تجربه کافی باید از رتبه‌بندی کنار گذاشته شوند. به علاوه، مصححین بی‌تجربه، از نظر اقتصادی، مقرون به صرفه‌تر از مصححین با تجربه هستند، و برای تصمیم گیرنده‌ها هزینه کمتری جهت رتبه‌بندی دارند. بنابراین، به جای تخصیص بودجه هنگفت به مصححین با تجربه، تصمیم گیرنده‌ها از بودجه استفاده بهتری در جهت ایجاد برنامه‌های آموزشی بهتر داشتند (بیژنی[3]، ۲۰۱۸).

در پژوهش‌های داخلی اثرات ارزیابان بر نمره‌گذاری آنها از متون فارسی که توسط دانشجویان غیرفارسی زبان نوشته شده است با استفاده از مدل‌های چند وجهی راش مورد بررسی قرار گرفته است. نتایج حاکی از آن است که از ۵ ملاک ارزیابی استفاده شده توسط ارزیابان (شامل دستور، دایره واژگانی، انسجام معنایی، تنوع محتوا و پرورش موضوع و ظاهر متن) مصححان نسبت به «انسجام معنایی» سوگیری بیشتری نشان داده‌اند و علاوه بر این، نسبت به هر دو موضوع نگارش نیز سوگیری داشته‌اند. نتایج، حاکی از تفاوت‌های معنی‌دار بین سوگیری‌های دو مصحح است (اسفندیاری[4]، ۲۰۱۴). تاثیر نوع ارزیابی (خود ارزیاب‌ها، ارزیابان همتا و ارزیابان معلم[5]) بر نتایج حاصل از آن نیز بررسی شده است. نتایج حاکی از آن است که ارزیاب‌های معلم سخت‌گیرتر، و خودارزیاب‌ها سهل‌گیرتر هستند و ارزیابان داخل هر گروه تغییرپذیری زیادی دارند. به علاوه از ۱۵ معیار ارزیابی (شامل، «محتوا»، «بسط جمله کلیدی مقاله»، «ارتباط موضوع مقاله با محتوای مقاله»، «مقدمه»، انسجام»، «نتیجه گیری»، «سیر منطقی»، «تنوع کلمات»، «انتخاب واژه‌ها»، «اجزای کلام»، «تنوع ساختارهای دستوری»، «دستور»، «املای واژه‌ها»، «شکل ظاهری کلمات»، و «به کار گیری علائم سجاوندی») در معیارهای «انتخاب کلمات»، «تنوع کلمات» و «ارتباط موضوع مقاله با محتوای مقاله» نسبت به بقیه معیارهای ارزیابی سخت‌گیرانه‌تر بوده است؛ در صورتی که در ارتباط با «سیر منطقی» و «املای کلمات» سهل‌گیری وجود داشته است (اسفندیاری و مایفورد[6]، ۲۰۱۳).

روش پژوهش

این طرح از نظر روش‌شناسی جزء روش‌های کمی از نوع توصیفی-همبستگی محسوب می‌شود. ارزیابانی که در سازمان سنجش برای نمره‌گذاری آزمون‌های مختلف عملکردی از آنها برای نمره‌گذاری استفاده می‌شود جامعه این پژوهش هستند. در این پژوهش با استفاده از روش‌های کلاسیک و رویکرد چندوجهی راش رفتار رتبه‌دهی ارزیابان تحلیل و بررسی شد و براساس شاخص‌های به دست آمده در مورد سوگیری‌های مد نظر موجود در داده‌ها قضاوت گردید. چون داده‌های این پژوهش قبلا توسط

[1]. Knoch
[2]. Wolfe & McVay
[3]. Bijani
[4]. Esfandiari
[5]. Self-assessors, peer-assessors, and teacher assessors
[6]. Myford

سازمان سنجش آموزش کشور جمع‌آوری شده از این حیث جزو داده‌های ثانویه محسوب می‌شوند به همین دلیل در این پژوهش نیاز به ابزار جمع‌آوری داده‌ها نیست. در پژوهش فعلی داده‌های مربوط به رتبه‌دهی ارزیابان در آزمون‌های طراحی معماری سال-های ۱۳۹۶ (۵۴۳۷ نفر)، اسکیس معماری ۱۳۹۷ (۷۴۵۹ نفر)، طراحی صنعتی سراسری سال ۱۳۹۶ (۱۳۶۵نفر)، موسیقی سال ۱۳۹۷ (۵۶۹ نفر) و نمایش عروسکی سال ۱۳۹۷ (۹۷ نفر) در اختیار پژوهشگران قرار گرفت. برای تحلیل با مدل‌های چندوجهی راش از بسته immer (روبیتز و استاینفیلد[1]، ۲۰۱۸a و ۲۰۱۸b) و برای محاسبه شاخص‌های همسانی و توافق از بسته IRR (گامر، لمون،فلوس و سینگ[2]، ۲۰۱۲) در R (تیم هسته R، ۲۰۱۹) استفاده شد.

در بین آزمون‌های تحلیل شده، تنها آزمونی که شرایط تحلیل با مدل چندوجهی راش را داشت آزمون طراحی صنعتی بود. با این حال، به دلیل نامناسب بودن طیف صفر تا صد استفاده شده، تحلیل با مدل‌های چند وجهی راش برای این آزمون نیز با مشکل مواجه شد (چرا که در طیف صفر تا صد تعداد زیادی از رتبه‌ها یا اصلا مورد استفاده قرار نگرفته بودند و در عین حال دامنه رتبه‌ها در هیچ یک از متغیرها به صد نیز نمی‌رسید). به همین دلیل ابتدا رتبه‌های صفر تا صد آزمون طراحی صنعتی بر ۱۰ تقسیم شدند تا طیف رتبه‌ها به بازه صفر تا ۱۰ تبدیل شد. سپس برای از بین بردن بخش اعشاری نمره‌های حاصل، رتبه‌ها به نزدیک‌ترین عدد صحیح گرد شدند. به این ترتیب در نهایت طیف داده‌ها به بازه صفر تا ۹ منتقل گردید.

یافته‌ها

برای تحلیل از روش‌های مختلف توصیفی مثل فراوانی، درصد، میانگین، میانه، نما، انحراف معیار، کجی و کشیدگی استفاده شده است. سپس از شاخص‌های مختلف پایایی، برای بررسی همسانی و اجماع و نیز از مدل‌های چند وجهی راش برای پاسخ‌گویی به سوال‌های پژوهش استفاده شده است. همان طور که در جدول ۱ مشخص است بیشترین شرکت‌کنندگان مربوط به آزمون اسکیس معماری سال ۱۳۹۷ و کمترین تعداد شرکت کننده مربوط به آزمون نمایش عروسکی همین سال است. بر اساس شاخص‌های توصیفی در آزمون اسکیس معماری سال ۱۳۹۷، رتبه‌های ارزیابان گروه سوم بیشترین میانگین (۱۳/۶۰) و کمترین پراکندگی (۱۰/۹۰) را دارد، در حالی که رتبه‌های ارزیابان گروه دوم کمترین میانگین (۱۱/۴۹) و بیشترین پراکندگی (۱۲/۷۵) را دارا هستند.

با توجه به طیف صفر تا صد رتبه‌ها می‌توان گفت در کل رتبه‌های اختصاص داده شده ارزیابان به کارهای عملی داوطلبان پایین است، که با توجه رقابتی بودن آزمون دور از انتظار نیست. چون توزیع رتبه‌ها کشیدگی مثبت دارد، مقادیر میانه به عنوان شاخص گرایش مرکزی مناسب‌تر، نشان می‌دهند که رتبه‌های اختصاص داده شده به کارهای عملی بسیار پایین است. در شاخص‌های توصیفی آزمون طراحی صنعتی سال ۹۶ رتبه‌های ارزیابان در سوال ۸ بیشترین میانگین (۴/۵۱) و پراکندگی (۳/۵۸) را دارد، در حالی که سوال‌های ۵ و ۶ کمترین میانگین (به ترتیب ۰/۷۳ و ۰/۴۶) و پراکندگی (به ترتیب ۱/۲۴ و ۱/۳۰) را دارا می‌باشند. اگر چه اختلاف چندانی بین این میانگین‌ها وجود ندارد، ولی با توجه به طیف صفر تا ۷۰ می‌توان گفت در کل رتبه‌های اختصاص داده شده ارزیابان به داوطلبان پایین است، که این موضوع ممکن است به دلیل دشواری سوال یا جدیت ارزیابان باشد.

سوال ۲ دارای کجی و کشیدگی منفی (به ترتیب ۰/۱۹- و ۱/۱۳-) و سوال ۶ دارای کجی وکشیدگی مثبت (به ترتیب ۳/۳۴ و ۱۱/۹۱) بالایی است که حاکی از وضعیت خاص رتبه‌ها در این دو سوال دارد. در شاخص‌های توصیفی آزمون موسیقی سال ۱۳۹۷ رتبه‌های ارزیابان در بخش هارمونی دارای بیشترین میانگین (۲۱/۲۸) و پراکندگی (۲۶/۹۴) و در بخش کنترپوان دارای کمترین میانگین (۷/۵۴) و پراکندگی (۱۷/۸۶) است. در بخش کنترپوان و شناخت موسیقی، میانه و نما هر دو صفر است که نشان دهنده این است که ۵۰٪ نمره‌ها بالاتر و۵۰٪ پایین‌تر از صفر قرار دارند. شاخص‌های توصیفی آزمون نمایش عروسکی سال ۱۳۹۷ نشان می‌دهدکه میانگین رتبه‌های داوطلبان در هر دو بخش نمایشنامه‌نویسی

[1]. Robitzsch & Steinfeld
[2]. Gamer, Lemon, Fellows & Singh

عروسکی (۱۶/۴۶) و نقد و تحلیل نمایش (۱۶/۳۵)، تقریبا یکسان است ولی پراکندگی رتبه‌ها در بخش نقد و تحلیل نمایش (۲۳/۸۵ در برابر ۱۷/۶۵ برای بخش نمایشنامه نویسی عروسکی) بیشتر است، و نما در هر دو بخش صفر است. شاخص‌های توصیفی آزمون طراحی معماری سال ۱۳۹۶ نشان می‌دهدکه رتبه‌های ارزیابان در بخش پلَن دارای بیشترین میانگین (۱۶/۵۲) و در بخش سکشن دارای کمترین میانگین (۸/۲۴) و بیشترین پراکندگی (۱۱/۴۱) است. رتبه‌های ارزیابان در بخش پرسپکتیو کمترین پراکندگی (۸/۶۷) را دارد. با توجه به طیف صفر تا ۱۰۰ رتبه‌ها می‌توان گفت رتبه‌های اختصاص داده شده داوران به کارهای عملی داوطلبان در سطح پایین قرار دارد، که با توجه به رقابتی بودن آزمون و یا جدیت ارزیابان در رتبه‌دهی دور از انتظار نیست.

جدول ۱: آماره‌های توصیفی آزمون‌های مختلف عملکردی در پژوهش حاضر

آزمون	متغیر	میانگین	میانه	نما	انحراف معیار	کجی	کشیدگی	کمینه	بیشینه
اسکیس معماری	گروه اول	۱۲/۱۱	۸	۵	۱۲	۱/۸۲(۰/۰۲۸)	۳/۲٤(۰/۰۵۷)	۰	۷۸
۱۳۹۷ (۷٤۵۹ نفر)	گروه دوم	۱۱/٤۹	٦	۳	۱۲/۷۵	۲/۲٤(۰/۰۲۸)	۵/٦۰(۰/۰۵۷)	۰	۸٦
	گروه سوم	۱۳/٦۰	۱۰	۱۰	۱۰/۹۰	۱/۵۲(۰/۰۲۸)	۲/۵۲(۰/۰۵۷)	۰	۸۰
	مجموع	۷۳/۱۹	۲۵	۱۸	۳۳/۵۱	۱/۸۱(۰/۰۲۸)	۳/۲۱(۰/۰۵۷)	۰	۲۲۸
	میانگین	۱۲/٤۰	۸/۳۳	٦	۱۱/۱۷	۱/۸۱(۰/۰۲۸)	۳/۲۱(۰/۰۵۷)	۰	۷٦
طراحی صنعتی	سوال۱	۱/٦۹	۱/۵۰	۱	۱/۲۱	۰/۷۲(۰/۰۷)	۰/۱۰(۰/۱۳)	۰/۰	٦
۱۳۹٦ (۱۳٦۵ نفر)	سوال۲	۳	۳	۳	۱/۸٦	-۰/۱۹(۰/۰۷)	-۱/۱۳(۰/۱۳)	۰/۰	٦
	سوال۳	۱/۷۸	۱/۵۰	۰/۰	۱/٤۲	۰/۷۳(۰/۰۷)	-/۰۲(۰/۱۳)	۰/۰	٦
	سوال٤	۲/۳۰	۲	۰/۰	۱/۸٦	۰/٤۵(۰/۰۷)	۰/٦۷(۰/۱۳)	۰/۰	۸
	سوال۵	۰/۷۳	۰/۰	۰/۰	۱/۲٤	۲/۳۸(۰/۰۷)	۷/۱۹(۰/۱۳)	۰/۰	۸
	سوال٦	۰/٤٦	۰/۰	۰/۰	۱/۳۰	۳/۳٤(۰/۰۷)	۱۱/۹۱(۰/۱۳)	۰/۰	۱۰
	سوال۷	۳/۱۰	۳	۲	۲/۲۳	۰/٦٤(۰/۰۷)	۰/۰۹(۰/۱۳)	۰/۰	۱۲
	سوال ۸	٤/۵۱	٤	۰/۰	۳/۵۸	۰/۳۹(۰/۰۷)	-۰/۸٦(۰/۱۳)	۰/۰	۱٤
	جمع نمرات	۱۷/٦۲	۱٦/۵۰	۱۳/۷۵	۹/٤۷	۰/۷۰(۰/۰۷)	۱/۰٤(۰/۱۳)	۰/۰	۵۳/۵
موسیقی ۱۳۹۷	هارمونی	۲۱/۲۸	٦	۰/۰	۲٦/۹٤	۱/۰۵(۰/۱۳)	۰/۱۹(۰/۲۰)	۰/۰	۹۷
(۵٦۹ نفر)	کنترپوان	۷/۵٤	۰/۰	۰/۰	۱۷/۸٦	۲/۵٦(۰/۱۳)	۵/۹۸(۰/۲۰)	۰/۰	۹۰
	شناخت موسیقی ایرانی	۱۳/٦۲	۰/۰	۰/۰	۲۳/۰۷	۱/۸۵(۰/۱۳)	۱/٦۲(۰/۲۰)	۰/۰	۱۰۰
	نمایشنامه نویسی	۱٦/٤٦	۱۲	۰	۱۷/٦۵	۱/۰۹(۰/۲۵)	۰/۷۲(۰/٤۹)	۰	۷٤
نمایش عروسکی	عروسکی								
۱۳۹۷ (۹۷ نفر)	نقد و تحلیل نمایش	۱٦/۳۵	۳	۰	۲۳/۸۵	۱/٦۱(۰/۲۵)	۱/٦۹(۰/٤۹)	۰	۹۰
طراحی معماری	ارزیاب ۱	۲۰/۲۷	۱۷	۱۵	۱۳/۹۰	۱/۷۰(۰/۰۳)	۳/۳٤(۰/۰۷)	۰	۹۲
۱۳۹٦ (۵٤۳۷ نفر)	ارزیاب ۲	۱۷/۱۸	۱۷	۱۸	۸/٤٤	۰/۵٤(۰/۰۳)	۱/۰۳(۰/۰۷)	۰	٦۰
بخش پلن	ارزیاب ۳	۱۲/۱۱	۸	۵	۱۱/۱۷	۱/۸٤(۰/۰۳)	٤/٤۱(۰/۰۷)	۰	۸۰
	کل	۱٦/۵۲	۱٤	۱۳/۳۳	۹/۹۲	۱/۲٤(۰/۰۳)	۲/۰۷(۰/۰۷)	۰	٦۸/۳۳
بخش پرسپکتیو	ارزیاب ۱	٤۵/۱۰	۸	۷	٦۷/۷	۵۰/۱(۰/۰۳)	۲۷/۳(۰/۰۷)	۰	۵٦
	ارزیاب ۲	۸۹/۷	۷	۱۰	٦٤/۵	۲٤/۱(۰/۰۳)	٤۱/۵(۰/۰۷)	۰	۸۰
	ارزیاب ۳	٤۵/۱۷	۱۵	۱۰	۱٤/۱۲	۸۸(۰/۰۳)	٤۵(۰/۰۷)	۰	٦٦
	کل	۱۱/۹۹	۱۱	۸/٦۷	۷/۲۰	۷۱(۰/۰۳)	۲۹(۰/۰۷)	۰	٤۳/٦۷
بخش سکشن	ارزیاب ۱	۳۱/٦	۲	۰	۲٤/۱۱	۹۹/۲(۰/۰۳)	۹۵/۱۰(۰/۰۷)	۰	۸۵

ارزیاب ۲	۵۱/٦	۲	°	٤٦/۹	٤٤/۲(۰/۰۳)	۲٤/۸(۰/۰۷)	°	۸۰
ارزیاب ۳	۷۸/۱۱	۵	°	۹۹/۱٤	٦۰/۱(۰/۰۳)	۲۷/۲(۰/۰۷)	°	۸۰
کل	۸/۲٤	۳/۳۳	°	۱۱/٤۱	۲/۱۲(۰/۰۳)	۵/۲۹(۰/۰۷)	°	۸۱/٦۷

*تعداد کل افراد ۱۴۹۲۷ نفر

در جدول ۲ منظور از همسانی، همبستگی بین رتبه‌های گروهی یا فردی ارزیابان در آزمون‌های اسکیس معماری و بخش‌های مختلف آزمون طراحی معماری است. شاخص استفاده شده در این حالت همبستگی پیرسون و تاو b کندال است. در صورتی‌که در آزمونهای طراحی صنعتی، موسیقی و نمایس عروسکی از شاخص آلفای بین سوال‌های آزمون برای نشان دادن همسانی درونی رتبه‌های ارزیابان به سوال‌ها استفاده شده است. باتوجه به نتایج جدول ۲ در آزمون اسکیس معماری، وقتی براساس همبستگی پیرسون در مورد همسانی بین ارزیابان قضاوت کنیم مقادیر همسانی قابل قبول است. ولی وقتی براساس تاو b کندال در مورد همسانی نمره‌گذاری ارزیابان قضاوت کنیم میزان همسانی چندان قابل قبول نیست، چرا که ضریب تاو کندال وجود گره در رتبه‌بندی‌ها را لحاظ می‌کند به همین دلیل مقدار همسانی کاهش می‌یابد. بر این اساس اگرچه میزان همسانی بین رتبه‌بندی ارزیابان تقریبا قابل قبول است ولی میزان توافق بین سه ارزیاب در آزمون اسکیس معماری به هیچ وجه رضایت بخش نیست (جدول ۲، ستون دوم از سمت چپ، کاپای کوهن بدون وزن و به خصوص وزن‌دار). در آزمون طراحی صنعتی نوع طرح استفاده شده به این صورت است که ۴ داور هشت سوال را ارزیابی کرده‌اند، ولی هر کدام از این ۴ داور فقط ۲ سوال از ۸ سوال را ارزیابی می‌کنند. به این ترتیب داور اول فقط سوال ۱ و ۲، داور دوم فقط سوال ۳ و ۴، داور سوم فقط سوال ۵ و ۶ و داور چهارم فقط سوال ۷ و ۸ را ارزیابی می‌کند. نمره‌گذاری سه سوال اول در طیف صفر تا ۶، نمره‌گذاری سوال چهار و پنج در طیف صفرتا ۸، نمره‌گذاری سوال ۶ در طیف صفر تا ۱۰ و نمره‌گذاری سوال ۷ در طیف صفر تا ۱۲، و نمره‌گذاری سوال ۸ در طیف صفر تا ۱۴ انجام می‌شود. با توجه به این که هر ارزیاب فقط دو سوال را ارزیابی می‌کند امکان محاسبه شاخص‌های محاسبه شده برای آزمون اسکیس در اینجا وجود ندارد.

جدول ۲: شاخص‌های همسانی (پیرسون، تاو b کندال و آلفا) همراه با پایایی اجماع (توافق) برای آزمون‌های عملکردی مختلف

آزمون	مقایسه ارزیاب	همبستگی پیرسون (تاو b کندال)	دو ارزیاب			چند ارزیاب	
			کاپای کوهن			کاپای فلیس	کاپای لایت
			بدون‌وزن	وزن-دار	عدم-توافق		
اسکیس معماری ۱۳۹۷	۱-۲	۰/۸٦(۰/٦۳)	۰/۰۸	۰/۲۰	۰/۱۵	۰/۰۵	۰/۰٦
	۱-۳	۰/۸۰(۰/۵۸)	۰/۰۵	۰/۱۱	۰/۱۰		
	۲-۳	۰/۷۹(۰/٦۱)	۰/۰٤	۰/۰۸	۰/۰٦		
طراحی معماری۱۳۹٦ (بخش پلان)	۱-۲	۰/٦۹(۰/۵٦)	۰/۰۷	۰/۳۵	۰/٤۱	۰/۰۵	۰/۰٦
	۱-۳	۰/٦۵(۰/۵۳)	۰/۰٤	۰/۰۲	-۰/۰۰۱		
	۲-۳	۰/۷۰(۰/۵۵)	۰/۰۵	۰/۰۸	۰/۰۷		
طراحی معماری۱۳۹٦ (بخش پرسپکتیو)	۱-۲	۰/۵۰(۰/٤۲)	۰/۰٦	۰/۱۲	۰/۱۵	۰/۰۵	۰/۰۵
	۱-۳	۰/٦٤(۰/۵۷)	۰/۰٦	۰/۰٤	۰/۳۸		
	۲-۳	۰/٤۱(۰/۳۹)	۰/۰۳	۰/۰۰۷	۰/۰۲		
طراحی معماری۱۳۹٦ (بخش سکشن)	۱-۲	۰/۸۵(۰/۸۱)	۰/۳۳	۰/۳۵	۰/٤۰	۰/۲۸	۰/۲۸
	۱-۳	۰/۸۹(۰/۸٦)	۰/۲٦	۰/۱۹	۰/۲۱		

	۲-۳	۰/۸۸(۰/۸۱)	۰/۲۶	۰/۲٤	۰/۲۹	
طراحی صنعتی ۱۳۹٦		۰/۷۳*			۰/۰۳	۰/۰۵
موسیقی ۱۳۹۷		۰/۳٤*			۰/۰۲	۰/۰٤
نمایش عروسکی ۱۳۹۷		۰/٤۷*			۰/۰۸	۰/۰۸

*آلفا فقط برای سوال‌های آزمون‌های طراحی صنعتی ۱۳۹۶، موسیقی۱۳۹۷ و نمایشنامه نویسی ۱۳۹۷ قابل محاسبه بود.

با توجه به داده‌های موجود برای این آزمون اگر فرض کنیم همه سوال‌ها یک چیز را اندازه‌گیری می‌کنند می‌توان شاخص آلفا، کاپای لایت، کاپای فلیز (و انواع ضرایب توافق درون طبقه‌ای که نتایج آن در ادامه ارایه شده است) را محاسبه کرد. میزان آلفا به عنوان شاخص همسانی درونی بین ارزیابان ۰/۷۳ است که یک مقدار مرزی است. این در صورتی است که ضرایب توافق کاپای فلیز و لایت بسیار پایین بوده و اجماع ضعیف را نشان می‌دهند. توجه کنید که با توجه به شرایط داده‌ها باید این شاخص‌ها را با احتیاط تفسیرکرد. در آزمون موسیقی، که دارای سه بخش هارمونی،کنترپوان و شناخت موسیقی است، در بخش هارمونی دو داور دو سوال را ارزیابی کرده‌اند. بنابر این هر کدام از دو داور هر دو سوال را مورد بررسی قرار داده‌اند. در بخش کنترپوان دو سوال توسط یک داور بررسی شده است و در بخش شناخت موسیقی سه سوال توسط دو داور مورد ارزیابی قرار گرفته است، بنابر این هر کدام از دو داور هر سه سوال را ارزیابی نموده‌اند. در اینجا نیز فرض می‌شود که نمره‌های بخش‌های مختلف نشان دهنده یک خصیصه هستند و همه داوران یک چیز را ارزیابی کرده‌اند و شاخص‌های مختلف همسانی (در اینجا آلفا)، توافق و همبستگی درون طبقه‌ای (که در ادامه گزارش شده‌اند) بر همین اساس محاسبه شده‌اند. مقادیر جدول ۲ برای آزمون موسیقی حاکی از همسانی و اجماع بسیار پایین است، که با توجه به وضعیت داده‌هایی که در اختیار پژوهشگر قرار گرفت دور از انتظار نیست.

در آزمون نمایش عروسکی که دارای دو بخش نمایشنامه نویسی و نقد و تحلیل نمایش است. هر کدام از داوران هر دو بخش را مورد ارزیابی قرار می‌دهند. باز هم با این فرض که نمره‌های بخش‌های مختلف نشان دهنده یک خصیصه هستند و همه داوران یک چیز را ارزیابی کرده‌اند، شاخص‌های مختلف همسانی (در اینجا آلفا)، توافق و همبستگی درون طبقه‌ای محاسبه شده‌اند (جدول ۲). مقادیر جدول ۲ برای آزمون نمایش عروسکی حاکی از همسانی و اجماع بسیار پایین است، که با توجه به وضعیت داده‌هایی که باز در اختیار پژوهشگر قرار گرفت دور از انتظار نیست. در آزمون طراحی معماری بر اساس نتایج جدول ۲ میزان همسانی بین رتبه‌بندی ارزیابان در بخش پلان و پرسپکتیو پایین است و در بخش سکشن تقریبا قابل قبول است. میزان توافق بین ارزیابان در هر سه بخش آزمون طراحی معماری در جدول ۲ اصلا رضایت بخش نیست.

جدول ۳: ضرایب همبستگی درون طبقه‌ای (ICC) برای رتبه‌های ارزیابان در آزمون‌های عملکردی مختلف

آزمون	مدل	ثبات		توافق	
		براساس رتبه هر فرد	براساس میانگین رتبه	براساس رتبه هر فرد	براساس میانگین رتبه
اسکیس ۱۳۹۷	یک طرفه	۰/۸۱(۰/۸۰-۰/۸۲)	۰/۹۳(۰/۹۲-۰/۹۳)	۰/۸۱(۰/۸۰-۰/۸۲)	۰/۹۳(۰/۹۲-۰/۹۳)
	دو طرفه	۰/۸۲(۰/۸۱-۰/۸۲)	۰/۹۳(۰/۹۳-۰/۹۳)	۰/۸۱(۰/۷۹-۰/۸۲)	۰/۹۳(۰/۹۲-۰/۹۳)
طراحی صنعتی ۱۳۹٦	یک طرفه	۰/۱٤(۰/۱۲-۰/۱٦)	۰/۵٦(۰/۵۳-۰/٦۰)	۰/۱٤(۰/۱۲-۰/۱٦)	۰/۵٦(۰/۵۳-۰/٦۰)
	دو طرفه	۰/۲۵(۰/۲۳-۰/۲۷)	۰/۷۳(۰/۷۱-۰/۷۵)	۰/۱۷(۰/۱۱-۰/۲۳)	۰/٦۲(۰/٤٦-۰/۷۳)
موسیقی ۱۳۹۷	یک طرفه	۰/۱۱(۰/۰٦-۰/۱٦)	۰/۲۷(۰/۱٦-۰/۳۷)	۰/۱۱(۰/۰٦-۰/۱٦)	۰/۲۷(۰/۱٦-۰/۳۷)
	دو طرفه	۰/۱۵(۰/۱۰-۰/۲۰)	۰/۳٤(۰/۲٤-۰/٤۳)	۰/۱۳(۰/۰۸-۰/۱۹)	۰/۳۲(۰/۱۹-۰/٤۲)
نمایش عروسکی ۱۳۹۷	یک طرفه	۰/۱۲(۰/۰٦-۰/٤۸)	۰/٤۸(۰/۲۲-۰/٦۵)	۰/۳۱(۰/۱۲-۰/٤۸)	۰/٤۸(۰/۲۲-۰/٦۵)
	دو طرفه	۰/۳۱(۰/۱۲-۰/٤۸)	۰/٤۷(۰/۲۱-۰/٦۵)	۰/۳۱(۰/۱۲-۰/٤۸)	۰/٤۷(۰/۲۱-۰/٦۵)
طراحی معماری ۱۳۹٦ (بخش پلان)	یک طرفه	۰/۵۵(۰/۵۳-۰/۵٦)	۰/۷۸(۰/۷۷-۰/۷۹)	۰/۵۵(۰/۵۳-۰/۵٦)	۰/۷۸(۰/۷۷-۰/۷۹)
	دو طرفه	۰/٦٤(۰/٦۳-۰/٦۵)	۰/۸٤(۰/۸۳-۰/۸۵)	۰/۵٦(۰/٤۱-۰/٦۷)	۰/۸۰(۰/٦٦-۰/۸۷)

طراحی معماری ۱۳۹٦ (بخش پرسپکتیو)	یک طرفه	۰/۳۱(۰/۳۰-۰/۳۳)	۰/۵۸(۰/۵٦-۰/٦۰)	۰/۳۱(۰/۳۰-۰/۳۳)	۰/۵۸(۰/۵٦-۰/٦۰)
	دو طرفه	۰/٤۸(۰/٤۷-۰/۵۰)	۰/۷٤(۰/۷۲-۰/۷۵)	۰/۳۷(۰/۱۸-۰/۵۲)	۰/٦٤(۰/۳۱-۰/۷۸)
طراحی معماری ۱۳۹٦ (بخش سکشن)	یک طرفه	۰/۷۷(۰/۷٦-۰/۷۸)	۰/۹۱(۰/۹۱-۰/۹۱)	۰/۷۷(۰/۷٦-۰/۷۸)	۰/۹۱(۰/۹۱-۰/۹۱)
	دو طرفه	۰/۸۳(۰/۸۲-۰/۸٤)	۰/۹٤(۰/۹۳-۰/۹٤)	۰/۷٦(۰/٦٤-۰/۸۵)	۰/۹۱(۰/۸٤-۰/۹۵)

جدول ۳ انواع ضرایب همبستگی درون طبقه‌ای[1] (ICC) آزمون‌های مختلف در حالت‌های گوناگون را نشان می‌دهد. ستون دوم از سمت راست دارای دو سطح یک طرفه و دو طرفه است. سطح یک طرفه به معنی آن است که فقط آزمودنی‌ها به طور تصادفی از جامعه مورد نظر انتخاب شده‌اند. سطح دو طرفه به این معنی است که هم داوطلبان و هم ارزیابان از جوامع مورد نظر به طور تصادفی انتخاب شده‌اند. ستون سوم و چهارم از راست دارای دو سطح ثبات و توافق هستند. از این رو وقتی دنبال ثبات (همسانی) هستیم مقادیر ردیف ثبات را مد نظر قرار می‌دهیم و وقتی دنبال توافق با شیم مقادیر ردیف توافق مد نظر قرار می‌گیرد. زیر بخش‌های ستون سوم و چهارم از راست به واحد تحلیل اشاره دارد که می‌تواند هر یک از افراد یا میانگین رتبه چند ارزیاب باشد. در این تحلیل برای محاسبه ضرایب ICC، ترکیب مختلف این عامل‌ها لحاظ شده است.

بر اساس نتایج جدول ۳ در آزمون اسکیس معماری وقتی فقط داوطلبان و یا هم داوطلبان و هم ارزیابان تصادفی در نظر گرفته شوند و هدف برآورد ثبات رتبه‌بندی‌ها یا توافق بین ارزیابان باشد، مقادیر شاخص ICC تقریبا ۰/۸۰ است، که اگرچه این مقدار قابل قبول است ولی نشان می‌دهد که تغییر سطوح مختلف عوامل تاثیر چندانی بر شاخص‌های ICC ندارد. در مقابل وقتی واحد تحلیل میانگین رتبه‌بندی ارزیابان است میزان شاخص ICC تقریبا تا ۰/۹۳ افزایش می‌یابد که نشان در این حالت میزان شاخص‌های ICC افزایش یافته است. یعنی وقتی رتبه هر فرد به عنوان واحد تحلیل انتخاب می‌شود میزان ثبات و توافق تقریبا یکسان است، ولی وقتی واحد تحلیل میانگین رتبه‌های ارزیابان است میزان ثبات و توافق تا حدی افزایش می‌یابد. از این رو می‌توان گفت تنها عامل موثرکه باعث می‌شود هم ثبات و هم توافق افزایش یابد میانگین رتبه بندی‌ها است.

در آزمون طراحی صنعتی براساس ضرایب همبستگی درون طبقه‌ای در جدول ۳ در مدل یک طرفه و دو طرفه همیشه تحلیل میانگین رتبه‌ها بالاتر از تحلیل رتبه هر فرد است، که باز با توجه وضعیت خاص داده‌ها دور از انتظار نیست. به خصوص وقتی که در مدل دو طرفه میانگین رتبه‌ها در حالت ثبات (همسانی) تحلیل می‌شوند.

در آزمون موسیقی براساس ضرایب همبستگی درون طبقه‌ای جدول ۳ در تمام حالات مقادیر بسیار پایین است ولی باز هم نتایج تحلیل براساس میانگین رتبه‌ها بیشتر از تحلیل رتبه هر فرد است، و باز هم مقادیر مدل دو طرفه بیشتر از مقادیر مدل یک طرفه است.

در آزمون نمایش عروسکی نیز براساس ضرایب همبستگی درون طبقه‌ای جدول ۳ در تمام حالات مقادیر بسیار پایین است ولی باز هم تحلیل میانگین رتبه‌ها بیشتر از تحلیل رتبه هر فرد است، و باز هم مقادیر مدل دو طرفه بیشتر از مقادیر مدل یک طرفه است

در آزمون طراحی معماری نیز با توجه به ضرایب همبستگی درون طبقه‌ای در جدول ۳ می‌توان گفت در مدل یک طرفه و دو طرفه همیشه تحلیل مبتنی بر میانگین رتبه‌ها بالاتر از تحلیل مبتنی بر رتبه هر فرد است. در اینجا نیز وقتی رتبه هر فرد به عنوان واحد تحلیل انتخاب می‌شود میزان ثبات و توافق پایین، و وقتی واحد تحلیل میانگین رتبه‌های ارزیابان است میزان ثبات و توافق افزایش می‌یابد.

یافته‌های مبتنی بر مدل‌های چندوجهی

با توجه به این که فقط داده‌های آزمون طراحی صنعتی شرایط تحلیل با مدل‌های چندوجهی راش را داشت، داده-های سایر آزمون‌ها با این مدل‌ها تحلیل نشد. همان طور که در بخش روش پژوهش ذکر شد، به دلیل مشکلات

[1]. Intraclass correlation

مربوط به نمره‌گذاری در آزمون طراحی صنعتی، مقیاس نمره‌ها از صفر تا صد از طریق تقسیم بر ۱۰ به طیف صفر تا نه تبدیل شد.

بر اساس نتایج جدول ۴ در بخش پلن، داور اول از همه طیف رتبه‌های ۱ تا ۹ استفاده نموده است، ولی بیشتر رتبه‌های ۱ تا ۳ را در نظر گرفته است. داور دوم و سوم بیشتر به استفاده از طیف رتبه‌های صفر تا ۵ تمایل داشته‌اند. همچنین میانگین رتبه داور اول بیشتر و میانگین رتبه داور سوم کمتر است. در بخش پرسپکتیو داور سوم از طیف رتبه‌های صفر تا ۸ استفاده نموده است ولی بیشتر به رتبه‌های صفر تا ۴ تمایل داشته است.

جدول ٤: آماره‌های توصیفی آزمون طراحی معماری ۱۳۹٦ در مقیاس صفر تا ۹

بخش	rater	9	8	7	6	5	4	3	2	1	0	mean	sd	corr
پلان	۱	۰/۰	۰/۰۱	۰/۰۱	۰/۰۲	۰/۰٤	۰/۰٦	۰/۰۹	۰/۳۸	۰/۳۵	۰/۰۵	۲	۱/٤۳	۰/٦۹
	۲	NA	NA	NA	۰/۰	۰/۰	۰/۰٤	۰/۱۰	۰/۵۰	۰/۲۸	۰/۰۸	۱/۷	۰/۹۱	۰/٦۹
	۳	NA	۰/۰	۰/۰	۰/۰	۰/۰۱	۰/۰۵	۰/۰٤	۰/۲۱	۰/۳۲	۰/۳٦	۱/۱	۱/۲۳	۰/٦۸
پرسپکتیو	۱	NA	NA	NA	۰/۰	۰/۰	۰/۰۱	۰/۰۳	۰/۱۹	۰/۵۱	۰/۲۵	۱/۰۵	۰/۸۵	۰/٦٤
	۲	NA	NA	NA	۰/۰	۰/۰	۰/۰	۰/۰	۰/۱۳	۰/٤۵	۰/٤۲	۰/۷۲	۰/۷۰	۰/٤٦
	۳	NA	NA	NA	۰/۱	۰/۰۳	۰/۰۷	۰/۱۳	۰/۲۸	۰/۳۳	۰/۱۵	۱/۷۳	۱/۲۷	۰/٦۳
سکشن	۱	NA	NA	NA	۰/۰	۰/۰	۰/۰۱	۰/۰۳	۰/۱۹	۰/۵۱	۰/۲۵	۱/۰۵	۰/۸۵	۰/٦٤
	۲	NA	NA	NA	۰/۰	۰/۰	۰/۰	۰/۰	۰/۱۳	۰/٤۵	۰/٤۲	۰/۷۲	۰/۷۰	۰/٦٦
	۳	NA	NA	NA	۰/۰۱	۰/۰۳	۰/۰۷	۰/۱۳	۰/۲۸	۰/۳۳	۰/۱۵	۱/۷۳	۱/۲۷	۰/٦۳

NA=not available

داور اول رتبه‌های ۱ تا ۳ را در نظر گرفته و داور دوم فقط طیف رتبه‌های صفر تا ۲ را استفاده نموده است. میانگین رتبه‌های بخش پرسپکتیو در داورسوم (۱/۷۳) بیشترین و در داور دوم (۰/۷۲) کمترین است. همچنین در بخش سکشن، داور سوم طیف رتبه‌های صفر تا ۸ را استفاده نموده، ولی بیشتر رتبه‌های صفر تا ۴ را در نظر گرفته است. داور اول فقط به استفاده از رتبه‌های صفر تا ۳ تمایل داشته است. و داور دوم فقط طیف رتبه‌های صفر تا ۲ را در نظر گرفته است. در بخش سکشن رتبه‌های داور سوم بیشترین میانگین (۱/۷۳) و رتبه‌های داور دوم کمترین میانگین (۰/۷۲) را داراست. در هر سه بخش بیشتر داوران به استفاده از رتبه‌های کرانه پایین تمایل داشته‌اند. در بخش پلان هر سه داور در نمره‌گذاری توافق دارند ولی در بخش پرسپکتیو داور سوم از نظر نمره‌گذاری با دیگر داوران توافق ندارد.

جدول ۵: مقایسه مدل‌های PCM و GPCM همراه با نتایج مدل LOCLCA برای تعداد طبقات پنهان مختلف در بخش پلن

Model	Deviance	par	AIC	BIC
PCM	۳۹۳۷۷,٦۸	۲٤	۳۹٤۲٦	۳۹۵۸٤
GPCM	۳۹۲٦٦,۳	۲٦	۳۹۳۱۸	۳۹٤۹۰
LOCLCA(3)	۳۹۱۱٤,۵٤	۲۲	۳۹۱۵۹	۳۹۳۰۳
LOCLCA(4)	۳۸۵٦۳,٦٦	۲٤	۳۸٦۱۲	۳۸۷۷۰
LOCLCA(5)	۳۸٤۲۰,۳۵	۲٦	۳۸٤۷۲	۳۸٦٤٤
LOCLCA(6)	۳۸٤۱۸,۳٦	۲۸	۳۸٤۷٤	۳۸٦۵۹
LOCLCA(7)	۳۸۳۸٦,۱٤	۳۰	**۳۸٤٤٦**	**۳۸٦٤٤**
LOCLCA(8)	**۳۸۳۸٤,۹۱**	۳۲	۳۸٤٤۹	۳۸٦٦۰
LOCLCA(9)	۳۸۳۸۵,٤٤	۳٤	۳۸٤۵۳	۳۸٦۷۷

بخش پلن آزمون طراحی معماری: مقایسه سه ارزیاب در بخش پلن براساس مدل‌های مختلف در جدول ۵ حاکی از آن است که با توجه به شاخص‌هایAIC و BIC، برازش مدل GPCM[1] بهتر از مدل PCM است، که نشان می‌دهد سه ارزیاب از نظر قدرت تشخیص و دقت با هم فرق دارند و آن‌طور که در مدل PCM مبتنی بر رویکرد چندوجهی راش فرض می شود، نمی‌توان آنها را از این جنبه یکسان در نظر گرفت. براساس نتایج حاصل از مدل‌های مختلف LOCLCA[2] جدول ۵ و با توجه به شاخص‌هایAIC و BIC برای هر یک از آنها، می‌توان گفت در بهترین حالت مدل حاوی هفت طبقه پنهان بهتر از مدل، PCM که توانایی را پیوسته فرض می‌کند، با داده‌ها برازش دارد. این هفت طبقه حاکی از طبقات پنهانی هستند که ارزیابان افراد را در آن‌ها قرار داده‌اند. البته خود شاخص Deviance از مدل حاوی هشت طبقه پنهان حمایت می‌کند.

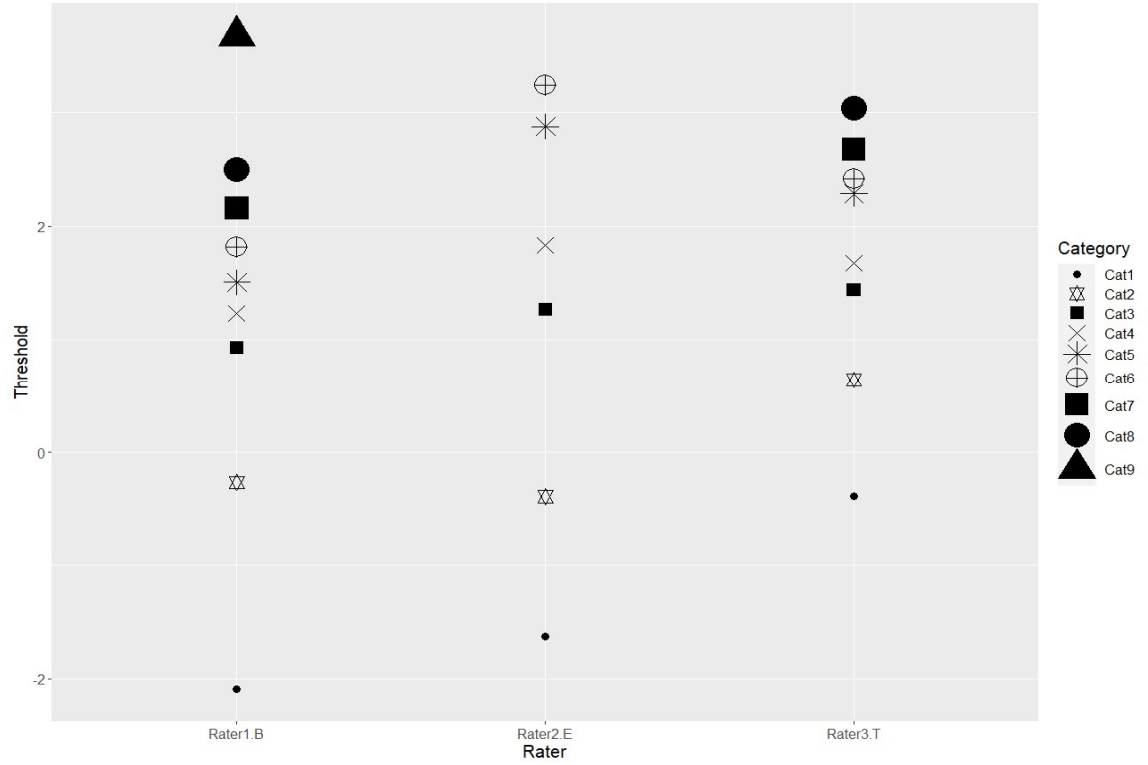
با توجه به پارامترهای بتای (Beta) سه ارزیاب در مدل GPCM که در جدول ۶ ارایه شده، می‌توان گفت در کل ارزیاب سوم سخت‌گیرترین فرد (۱/۷۲) و ارزیاب دوم سهل‌گیرترین فرد (۱/۲۰) است. این وضعیت به تفکیک هر یک از طبقات صفر تا ۹ نیز قابل مشاهده است (مقادیر زیر هر یک از ستون‌های Cat1 تا Cat9 برای هر یک از سه ارزیاب را مقایسه کنید. عدد کمتر نشان‌دهنده سهل‌گیری و عدد بزرگتر نشانه سخت‌گیری است. عبارت NA به عدم استفاده ارزیاب از رتبه مربوطه اشاره دارد). به علاوه براساس مقادیر ستون آلفا (alpha)، ارزیاب دوم دارای بیشترین قدرت تشخیص (۲/۷۳) و ارزیاب اول دارای کمترین قدرت تشخیص (۲/۱۰) است.

جدول ٦: آماره‌های مربوط به سه ارزیاب در آزمون طراحی معماری بخش پلان

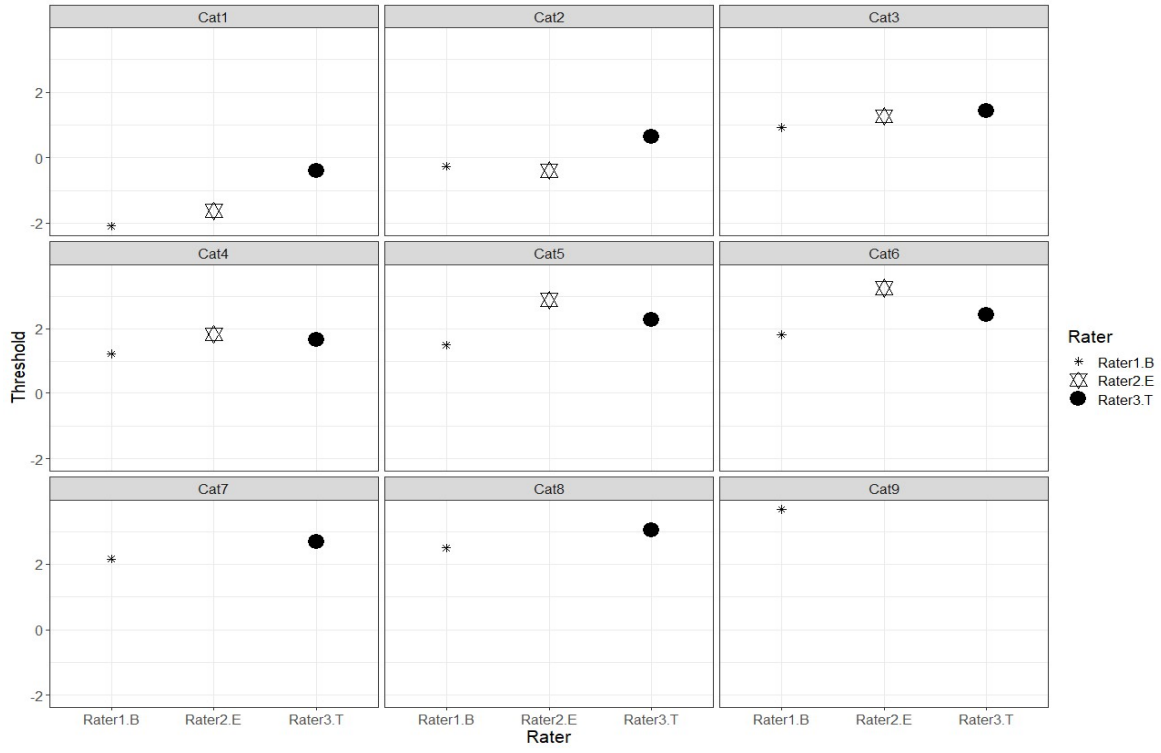
ارزیاب	alpha	Beta	Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7	Cat8	Cat9
۱	۲/۱۰	۱/۲۷	-۳/۳۵	-۰/۵۳	-۰/۱۲	-۰/۰٦	۰/۲۰	۰/۵۲	۰/۹٤	۱/۰۲	۲/۳۸
۲	۲/۷۳	۱/۲۰	-۲/۸۲	-۱/٦۰	۰/۱۵	۰/۵٦	۱/۸۳	۱/۸۸	NA	NA	NA
۳	۱/٦۵	۱/۷۲	-۲/۰۱	-۱/۱۰	۰/۱۷	-۰/۵٤	۱/۲۰	۰/۳٤	۰/۹۵	۰/۹۸	NA

براساس نتایج نمودار ۱ ارزیاب اول از تمام طبقات طیف نمره‌گذاری استفاده کرده، ولی ارزیاب سوم و به خصوص دوم از طبقات طیف بالای مقیاس نمره‌گذاری اصلا استفاده نکرده‌اند، که حاکی از سخت‌گیری آنها در استفاده از این طبقات است. ارزیاب دوم و به خصوص اول در استفاده از طبقات اول سهل‌گیر بوده‌اند. با توجه به اینکه تغییرپذیری در طبقات پایین بیشتر از طبقات بالا است می‌توان گفت توافق بین ارزیابان در طبقات بالا بیشتر از طبقات پایین است. هر چند که این توافق بسیار پایین است. به نظر می‌رسد نوعی گرایش به مرکز در ارزیاب دوم و به خصوص سوم وجود دارد. براساس نمودار ۲ سخت‌گیری داور سوم تقریبا در تمام طبقات مشهود است، به خصوص طبقات بالا که اصلا توسط وی و ارزیاب دوم استفاده نشده‌اند. ارزیاب اول از تمام طبقات نمره‌گذاری استفاده کرده است. بخش پر سپکتو آزمون طراحی معماری: براساس شاخص‌های برازش Deviance، AIC و BIC جدول ۷، باز هم برازش مدل GPCM بهتر از مدل PCM است که نشان می‌دهد، سه ارزیاب از نظر دقت یکسان نیستند. براساس شاخص‌های برازش Deviance، AIC و BIC مدل‌های مختلف LOCLCA می‌توان گفت در بهترین حالت مدل حاوی ۴ و یا ۵ طبقه پنهان بهتر از مدل PCM که توانایی را پیوسته فرض می‌کند با داده‌ها برازش دارد. این ۵ طبقه حاکی از طبقات پنهانی هستند که ارزیابان افراد را در آن‌ها قرار داده‌اند. به بیان دقیق‌تر، شاخص‌های برازش Deviance و AIC از مدل حاوی ۵ طبقه پنهان و شاخص BIC از مدل حاوی ۴ طبقه پنهان حمایت کرده است.

[1]. generalized partial credit model
[2]. located latent class Rasch models



نمودار ۱: مقایسه سه ارزیاب بخش پلن براساس ۹ طبقه. محور افقی از چپ به راست ارزیاب ۱ تا ۳ و محور عمودی دشواری هر یک از ۹ طبقه



نمودار ۲: مقایسه سه ارزیاب در بخش پلن براساس براساس ۹ طبقه مختلف رتبه‌بندی

جدول ۷: مقایسه مدل‌های PCM و GPCM همراه با نتایج مدل LOCLCA برای تعداد طبقات پنهان مختلف در بخش پرسپکتو

Model	Deviance	par	AIC	BIC
PCM	۳۶۷۱۶,۰۷	۲۲	۳۶۷۶۰	۳۶۹۰۵
GPCM	۳۶۳۹۶,۵	۲٤	۳٦٤٤٤	۳٦٦۰۳
LOCLCA(3)	۳٦٤٦۰,٤۲	۲۲	۳٦٥۰٤	۳٦٦٤۹
LOCLCA(4)	۳٦۳۹۹,٥۳	۲٤	۳٦٤٤۸	۳٦٦۰٦
LOCLCA(5)	۳٦۳۹۳,۷٤	۲٦	۳٦٤٤٦	۳٦٦۱۷
LOCLCA(6)	۳٦۳۹٤,۹۲	۲۸	۳٦٤٥۱	۳٦٦۳٥
LOCLCA(7)	۳٦۳۹۳,۷۹	۳۰	۳٦٤٥٤	۳٦٦٥۱

با توجه به پارامترهای بتای (Beta) سه ارزیاب براساس مدل GPCM در جدول ۸ می‌توان گفت در کل در بخش پرسپکتیو، ارزیاب دوم سخت‌گیرترین فرد (۲/۴۳) و ارزیاب سوم سهل‌گیرترین فرد (۱/۴۳) است. این وضعیت به تفکیک هر یک از طبقات صفر تا ۸ نیز قابل مشاهده است مقادیر زیر هر یک از ستون‌های Cat1 تا Cat8 برای هر یک از سه ارزیاب را مقایسه کنید. عدد کمتر نشان‌دهنده سهل‌گیری و عدد بزرگتر نشانه سخت‌گیری است. عبارت NA به عدم استفاده ارزیاب از رتبه مربوطه اشاره دارد). به علاوه براساس مقادیر ستون آلفا (Alpha) ارزیاب اول دارای بیشترین قدرت تشخیص (۳/۲۸) و ارزیاب دوم (۱/۰۷) دارای کمترین قدرت تشخیص است.

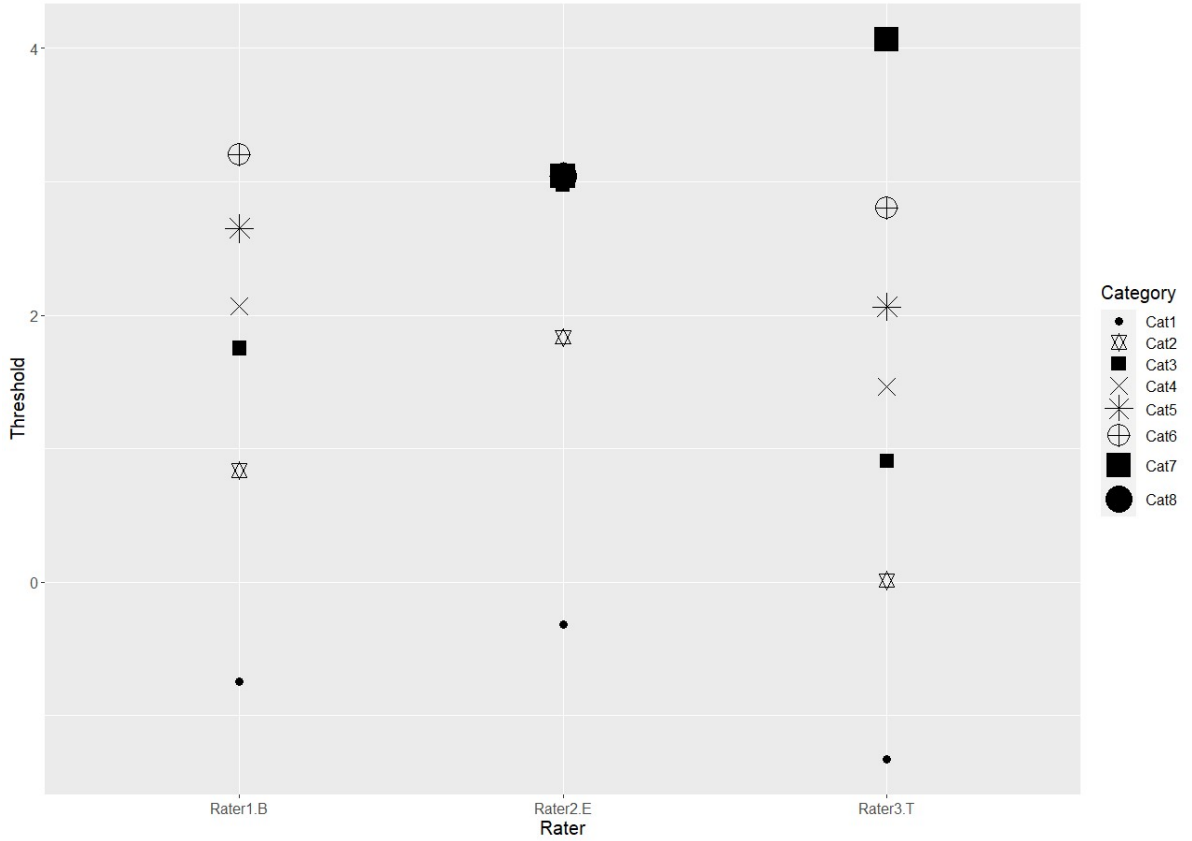
جدول ۸: آماره‌های مربوط به سه ارزیاب در آزمون طراحی معماری بخش پرسپکتیو

ارزیاب	Alpha	Beta	Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7	Cat8
۱	۳/۲۸	۱/٦۳	-۲/۳۷	-۰/۷۸	۰/۲۳	۰/۳٥	۱/۰٤	۱/٥۲	NA	NA
۲	۱/۰۷	۲/٤۳	-۲/٦٥	-۰/٦۳	۱/۹٤	۲/۰٥	٦/٥۸	۰/۸۸	-۱/۱۷	-۷/۰۱
۳	۱/۷۰	۱/٤۳	-۲/۷۰	-۱/۳۷	-۰/٤۲	۰/۰۱	۰/٦۱	۱/۳۰	NA	۲/٥۸

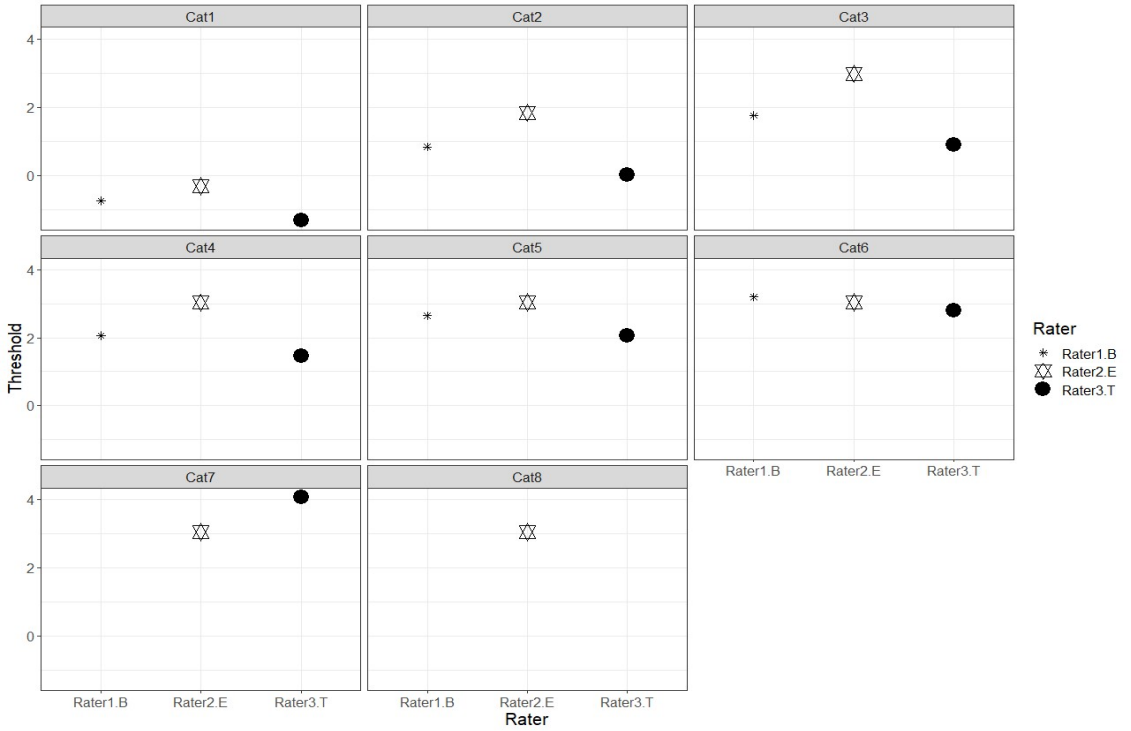
براساس نتایج نمودار۳ هر سه ارزیاب از طبقه ۹ مقیاس نمره‌گذاری اصلا استفاده نکرده‌اند که حاکی از سخت‌گیری آنها در استفاده از این طبقه است. ارزیاب اول و خصوصا سوم در استفاده از طبقات اول سهل‌گیر بوده‌اند.

با توجه به اینکه تغییرپذیری در طبقات پایین بیشتر از طبقات بالا است می‌توان گفت توافق بین ارزیابان در طبقات بالا بیشتر از طبقات پایین است. هر چند که این توافق بسیار پایین است. به نظر می‌رسد نوعی گرایش به مرکز در ارزیاب دوم و به خصوص اول وجود دارد. بخش سکشن آزمون طراحی معماری: مقایسه سه ارزیاب با استفاده از شاخص‌های برازش Deviance، AIC و BIC در بخش سکشن براساس مدل‌های GPCM و PCM در جدول ۹ حاکی از آن است که باز هم مطابق انتظار برازش مدل GPCM بهتر از مدل PCM است، که نشان می‌دهد دقت سه ارزیاب با هم فرق دارد.

براساس نتایج حاصل از مدل‌های مختلف LOCLCA در جدول ۹ و با توجه به شاخص‌های برازش Deviance، AIC و BIC می‌توان گفت در بهترین حالت مدل حاوی هشت طبقه پنهان بهتر از مدل PCM که توانایی را پیوسته فرض می‌کند با داده‌ها برازش دارد. این هشت طبقه حاکی از طبقات پنهانی هستند که ارزیابان افراد را در آن‌ها قرار داده‌اند. براساس پارامترها بتای (Beta) مدل GPCM برای سه ارزیاب در جدول ۱۰ می‌توان گفت که در کل ارزیاب دوم سخت‌گیرترین فرد (۲/۰۷) و ارزیاب سوم سهل‌گیری‌ترین فرد (۱/۴۸) است.



نمودار ۳: مقایسه سه ارزیاب بخش پرسپکتیو در ۹ طبقه. محور افقی از چپ به راست ارزیاب ۱ تا ۳ و محور عمودی دشواری هر یک از ۹ طبقه



نمودار ٤: مقایسه سه ارزیاب در بخش پرسپکتیو براساس براساس ۹ طبقه مختلف رتبه‌بندی

جدول ۹: مقایسه مدل‌های PCM و GPCM همراه با نتایج مدل LOCLCA برای تعداد طبقات پنهان مختلف در بخش سکشن

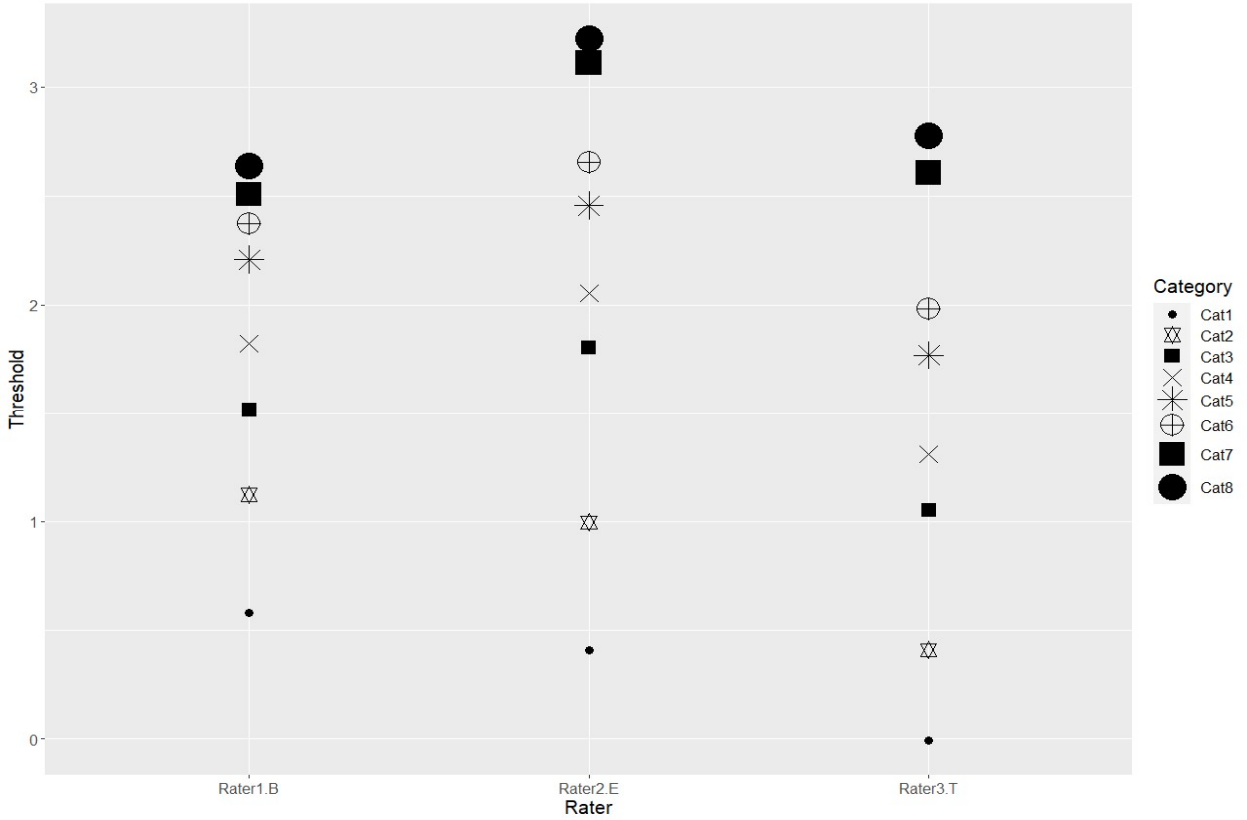
Model	Deviance	par	AIC	BIC
PCM	۲۷۷۳۳,۸۲	۲۵	۲۷۷۸٤	۲۷۹٤۸
GPCM	۲٦۷۹۷,٦٤	۲۷	۲٦۸۵۲	۲۷۰۲۹
LOCLCA(3)	۲۷٦٤۱,۹۵	۲۸	۲۷٦۹۸	۲۷۸۸۲
LOCLCA(4)	۲٦۹۳۹,۰۵	۳۰	۲٦۹۹۹	۲۷۱۹۷
LOCLCA(5)	۲٦۷۳۹,۵۳	۳۲	۲٦۸۰٤	۲۷۰۱٤
LOCLCA(6)	۲٦۷۳۸,۷٤	۳٤	۲٦۸۰۷	۲۷۰۳۱
LOCLCA(7)	۲٦٦٤۸,۵۵	۳٦	۲٦۷۲۱	۲٦۹۵۸
LOCLCA(8)	۲٦٦۱۷,۲۷	۳۸	۲٦٦۹۳	۲٦۹٤٤
LOCLCA(9)	۲٦٦۲٦,۲۳	٤۰	۲٦۷۰٦	۲٦۹۷۰

این وضعیت به تفکیک هر یک از طبقات صفر تا ۹ نیز قابل مشاهده است (مقادیر زیر هر یک از ستون‌های Cat1 تا Cat8 برای هر یک از سه ارزیاب را مقایسه کنید. عدد کمتر نشان‌دهنده سهل‌گیری و عدد بزرگتر نشانه سخت‌گیری است. عبارت NA به عدم استفاده ارزیاب از رتبه مربوطه اشاره دارد). به علاوه براساس مقادیر ستون آلفا (alpha) ارزیاب اول دارای بیشترین قدرت تشخیص (۵/۲۱) و ارزیاب دوم دارای کمترین قدرت تشخیص (۳/۸۷) است.

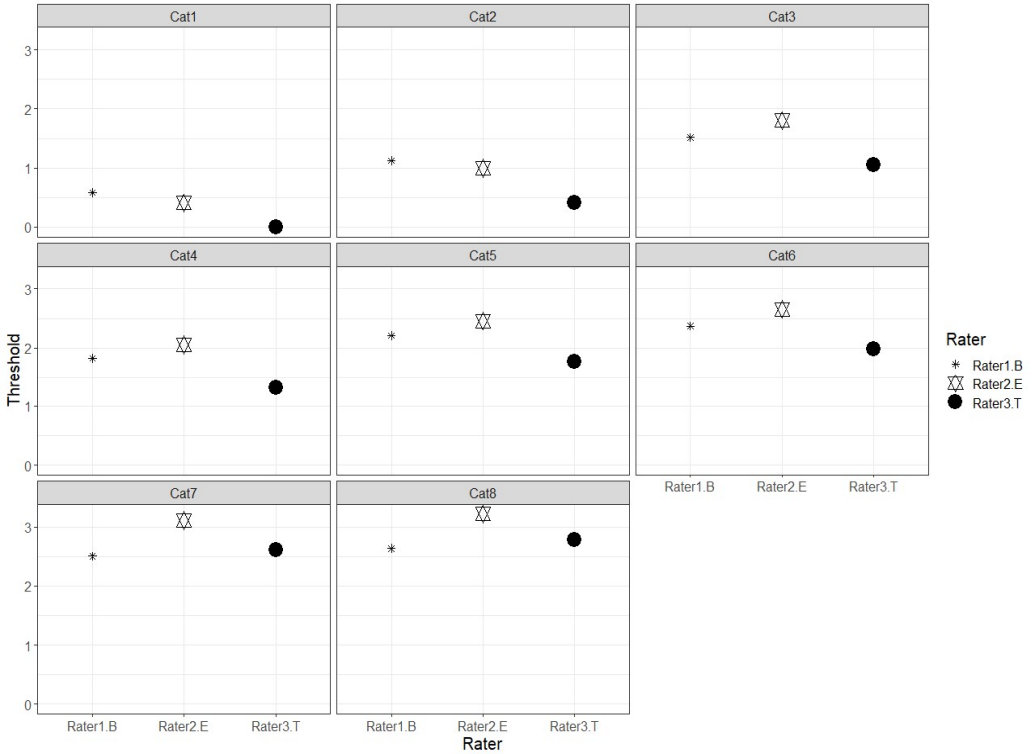
جدول ۱۰: آماره‌های مربوط به سه ارزیاب در آزمون طراحی معماری بخش سکشن

ارزیاب	alpha	Beta	Cat1	Cat2	Cat3	Cat4	Cat5	Cat6	Cat7	Cat8
۱	۵/۲۱	۱/۸۵	-۱/۲۵	-۰/۷۱	-۰/۳۱	-۰/۰٤	۰/٤۲	۰/۵۳	۰/٦۹	۰/٦۸
۲	۳/۸۷	۲/۰۷	-۱/٦۵	-۱/۱۱	-۰/۱۸	-۰/۱۱	۰/٤۷	۰/٤۵	۱/۲٦	۰/۸۷
۳	٤/۹٦	۱/٤۸	-۱/٦۷	-۱/۱۰	-۰/۳۸	-۰/۲۳	۰/۳۵	۰/٤۲	NA	۲/۵۸

براساس نتایج نمودار ۵ و ۶ هر سه ارزیاب از طبقه ۹ مقیاس نمره‌گذاری اصلا استفاده نکرده‌اند که حاکی از سخت‌گیری آنها در استفاده از این طبقات است. ارزیاب اول و خصوصا سوم در استفاده از طبقات اول سهل‌گیر بوده‌اند. با توجه به اینکه تغییرپذیری در طبقات پایین بیشتر از طبقات بالا است می‌توان گفت توافق بین ارزیابان در طبقات بالا بیشتر از طبقات پایین است. هر چند که این توافق بسیار پایین است. به نظر می‌رسد نوعی گرایش به مرکز در ارزیاب دوم و اول نیز وجود دارد.



نمودار ۵: مقایسه سه ارزیاب بخش سکشن براساس ۹ طبقه. محور افقی از چپ به راست ارزیاب ۱ تا ۳ و محور عمودی دشواری هر یک از ۹ طبقه



نمودار ۶: مقایسه سه ارزیاب در بخش سکشن براساس براساس ۹ طبقه مختلف رتبه‌بندی

بحث و نتیجه‌گیری

با توجه به بحث مطابقت مطلق (که با روش‌هایی مثل شاخص توافق دقیق و کاپای وزن‌دار کوهن ارزیابی می‌شود) و نسبی (که با روش‌هایی مثل ضریب همبستگی گشتاوری پیرسون و تاو b کندال ارزیابی می‌گردد) درجه‌بندی‌ها، می‌توان گفت در آزمون‌های اسکیس معماری شاخص‌های هم سانی مثل همبستگی پیر سون حاکی از هم سانی قابل قبول رتبه‌بندی ارزیابان ا ست، ولی برا ساس تاو b کندال میزان توافق مطلق چندان قابل قبول نیست، که با توجه به سرنو شت ساز بودن این آزمون نگران کننده ا ست. به بیان دقیق‌تر، بر ا ساس شاخص‌های توافق نتایج نشان داد که در آزمون‌های اسکیس معماری وقتی ارزیابان دوبه دو با هم مقایسه شوند یا وقتی هر سه ارزیاب با هم مقایسه شوند میزان توافق یا اجماع ب سیار پایین ا ست که با توجه به اهمیت این آزمون ا صلا قابل قبول نی ست. در آزمون طراحی معماری فقط در بخش سک شن همسانی قابل قبولی بین سه ارزیاب وجود دارد. در بخش‌های پلان و پرسپکتیو همسانی در حد متوسط و پایین است که قابل قبول نیست.

براساس شاخص‌های اجماع در بخش‌های پلان و پرسپکتیو توافق بین ارزیابان بسیار پایین است، در حالی که در بخش سکشن یک توافق نسبی ضعیف بین ارزیابان وجود دارد. این موضوع نشان می‌دهد که احتمالا بخش سکشن از نظر محتوایی با بخش-های پلان و پرسپکتیو متفاوت است. اثر جدیت به این معنی است که ارزیاب تمایل دارد پایین‌تر از نقطه میانی مقیاس اندازه-گیری، درجه‌بندی نماید. اثرات ارزیاب، از جمله جدیت، می‌تواند نمره‌های آزمون‌ها را تا حد زیادی تحت تاثیر قرار دهند و در نتیجه داوطلبان حتی اگر از توانایی بالایی هم برخوردار باشند قربانی این اثر ارزیاب می‌شوند. بهتر است برای کاهش اثرات جدیت و تاثیر آن بر رتبه‌دهی آزمون‌ها و مقابله با جدیت بیشتر ارزیابان مقیاس‌های درجه‌بندی طراحی شود، که در سمت مثبت آن چند طبقه مقیاس و در سمت منفی آن تعداد کمتری طبقه مقیاس وجود دارد؛ همچنین برای تعدیل نمره‌های ارزیابان در جدیت از روش‌های آماری استفاده شود (اکس ، ۲۰۱۱). ارزیابان در اثر گرایش به مرکز از طبقه وسط مقیاس اندازه‌گیری بیش از حد استفاده می‌کنند. پیامد گرایش به مرکز در نمره‌گذاری آزمون‌ها این است که درجه‌بندی‌ها در وسط مقیاس جمع شده و طیف موثر مقیاس کاهش یافته و باعث می‌شود قدرت تشخیص درجه‌بندی‌ها کاهش یابد. این موضوع موجب می‌شود هم پایایی و هم روایی کاهش یابد (دکاتیس[1]، ۱۹۷۷). مطابق انتظار در پژوهش حاضر اثر گرایش به مرکز در بین ارزیابان مشاهده نشد، بلکه نتایج بیشتر از سخت‌گیری آنها حکایت داشت. با توجه به نتایج حاصل از پژوهش حاضر می‌توان گفت ارزیابان می‌توانند نمرات فراگیران را تا حد زیادی تحت تاثیر قرار دهند. بنابر این نیاز است به ارزیابان قبل از تصحیح و نمره‌گذاری آزمون‌ها آموزش لازم و کافی داده شود تا آن‌ها کمترین میزان خطا را در زمان ارزیابی و نمره‌گذاری داشته باشند. این آموزش می‌تواند به روش‌های مختلف انجام شود. به علاوه لازم است طرح نمره‌گذاری استفاده شده در آزمون‌های عملکردی مورد بازبینی قرار بگیرد به طوری که امکان ارزیابی داوران از روی داده‌ها با مدل‌های مختلف فراهم شود. از سوی دیگر طیف نمره‌گذاری صفر تا ۱۰۰ یا صفر تا ۷۰ برای نمره‌گذاری آزمون‌های عملی سازمان سنجش مناسب نیست. به این دلیل که طیف استفاده شده آنقدر گسترده است که برخی از طبقات آن اصلا توسط ارزیابان استفاده نمی‌شود و در عین حال تصمیم‌گیری در خصوص اختصاص رتبه را برای ارزیاب با مشکل همراه می‌سازد. به علاوه در پیشینه ارزیابی آزمون‌های عملی سیستمی شبیه این طیف مشاهده نشد. به همین دلایل بهتر است طیف محدوتری مثل صفر تا ۱۰ برای این منظور استفاده شود. این طور به نظر می‌رسد که طیف صفر تا ۱۰۰ یا صفر تا ۷۰ استفاده شده توسط ارزیابان به هنگام ارزیابی کارهای عملی به عنوان نمره مورد استفاده قرار می‌گیرد نه رتبه‌ای که ارزیابان به کارهای افراد اختصاص می‌دهند، که البته جای بحث دارد. در کنار این موضوع باید توجه شود که مطابق نتایج حاصل از مدل‌های

[1]. DeCotiis

چند وجهی راش در بخش‌های مختلف آزمون طراحی معماری ارزیابان تمایل دارند داوطلبان را در بخش پلن در هفت تا هشت طبقه، در بخش پرسپکتو در چهار تا پنج و در بخش سکشن در هشت گروه طبقه‌بندی کنند. این یافته تاییدی بر عدم تناسب و ناهمخوانی طیف نمره‌گذاری استفاده شده با آنچه که ارزیابان در عمل انجام می‌دهند است. به علاوه لازم است استفاده از رویکردهای موجود در مدل‌های چند وجهی راش مورد توجه قرار گیرد، چرا که می‌تواند مقیاس‌های دقیق‌تر و کاربردی‌تری فراهم سازد تا ارزیابان بتوانند با استفاده از آن، نمرات دقیق‌تری به آزمون‌های عملی بدهند. چرا که آماره‌های موجود در رویکرد مدل‌های چند وجهی به ما می‌گویند که تا چه اندازه مقیاس‌های طراحی شده برای نمره‌گذاری دقیق است.

References

Andrich, D. (2005). *Rasch, Georg. Encyclopedia of Social Measurement*, 3, 299–306. Angeles, CA: Sage.

Bijani, H. (2018). Effectiveness of A Training Program on Oral Performance Assessment: The Analysis of Tasks Using the Multifaceted Rasch Analysis. *Journal of Modern Research in English Language Studies*, 5(4), 27-53.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY: Routledge.

DeCotiis, T. A. (1977). An analysis of the external validity and applied relevance of three rating formats. *Organizational Behavior and Human Performance*, 19, 247-266.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement*. Franfurt am Main: Peter Lang.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Engelhard Jr, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, 31(2), 93-112.

Esfandiari, R. (2014). A Many-Facet Rasch Measurement of Bias among Farsi-Native Speaking Raters toward Essays Written by Non-Native Speakers of Farsi. *Journal of Teaching Persian to Speakers of Other Languages*, 3(VOL.3,NO.3,(TOME 8)), 25-54.

Esfandiari, R. & Myford, C. M. (2013). Severity Differences Among Self-Assessors, Peer-Assessors, and Teacher Assessors Rating EFL Essays. Assessing Writing, 18(2): 111-131.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, NJ: Wiley.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *IRR: various coefficients of interrater reliability and agreement. 2012*. R package version 0.84, 1.

Hays, W. L. (1994). *Statistics* (5th ed.). Belmont, CA: Wadsworth.

Kempf, W. F. (1972). Probabilistische Modelle experimentalpsychologischer Versuchssituationen [Probabilistic models of designs in experimental psychology]. *Psychologische Beiträge*, 14, 16–37.

Kim, S. C., & Wilson, M. (2009). A comparative analysis of the ratings in performance assessment using generalizability theory and the many-facet Rasch model. *Journal of applied measurement*, 10(4), 408–423.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior: A longitudinal study. *Language Testing*, 28, 179–200.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (2006a). *Demarcating category intervals. Rasch Measurement Transactions*, 19, 1041–1043.

Linacre, J. M. (2014b). *A user's guide to FACETS: Rasch-model computer programs*. Chicago: Winsteps.com. Retrieved from http://www.winsteps.com/facets.

Linacre, J. M., & Wright, B. D. (1989). The length of a logit. *Rasch Measurement Transactions*, 3, 54–5.

Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measuremen*t, 3, 484–509.

Ludlow, L. H., & Haley, S. M. (1995). Rasch model logits: Interpretation, use, and transformation. *Educational and Psychological Measurement*, 55, 967–975.

Masters, G. N. (2010). The partial credit model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 109–122). New York, NY: Routledge.

Micko, H. C. (1970). Eine Verallgemeinerung des Meßmodells von Rasch mit einer Anwendung auf die Psychophysik der Reaktionen [A generalization of Rasch's measurement model with an application to the psychophysics of reactions]. *Psychologische Beiträge*, 12, 4–22.

Myers, J. L., Well, A. D., & Lorch, R. F. (2010). *Research design and statistical*.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of applied measurement*, 4(4), 386-422.

Ostini, R., & Nering, M. L. (2006). Polytomous item response theory models. Thousand Oaks, CA: Sage.

Penfield, R. D. (2014). An NCME instructional module on polytomous item response theory models. *Educational Measurement: Issues and Practice*, 33(1), 36–48.

Robitzsch, A., & Steinfeld, J. (2018a). *immer: Item response models for multiple ratings*. R package version 1.1-35.

Robitzsch, A., & Steinfeld, J. (2018b). Item response models for human ratings: Overview, estimation methods, and implementation in R. *Psychological Test and Assessment Modeling*, 60(1), 101-139.

Smith Jr, E. V., & Kulikowich, J. M. (2004). An application of generalizability theory and many-facet Rasch measurement using a complex problem-solving skills assessment. *Educational and Psychological Measurement*, 64(4), 617-639.

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29–49). Los

Tinsley, H. E. A., & Weiss, D. J. (2000). Interrater reliability and agreement. In Wolfe, E. W. (1997). *The relationship between essay reading style and scoring proficiency in a psychometric scoring system. Assessing Writing*, 4, 83–106.

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, 31(3), 31–37.