Educational Measurement and Evaluation Studies

National Organization for Educational Testing

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

Multi-Faceted Rasch Model in Practical Tests: Study Subject: Sorayesh Test

Seyyedeh Hoda Naji[1], Ali Moghadamzadeh[2], Balal Izanloo[3], Ebrahim Khodaie[4]

1. Ph.D student. Department of Curriculum Planning and Educational Methods, Faculty of Psychology and Education, University of Tehran, Tehran, Iran. Email: s.hodanaji@ut.ac.ir
2.Associate professor, Department of Curriculum Planning and Educational Methods, Faculty of Psychology and Education, University of Tehran, Tehran, Iran; (Corresponding Author), Email: amoghadamzadeh@ut.ac.ir
3 Associate Professor, University of Kharazmi, Faculty of Psychology and Education , Karaj, Iran. Email: izan.b@khu.ac.ir
4. associate professor. Department of Curriculum Planning and Educational Methods, Faculty of Psychology and Education, University of Tehran, Tehran, Iran. Email: khodaie@ut.ac.ir

Article Info

Article Type:
Research Article

Received:2024/03/24
Received in revised form:2024/08/10
Accepted:2024/09/05
Published online 2024/10/06

ABSTRACT

Objective: The purpose of this research is to analyze the data obtained from the practical tests conducted by National Organization of Educational Testing, using the Multi-Faceted Rasch model and comparing it with the results of classical analysis methods.

Methods: The research method is quantitative a type of descriptive-analysis method. The participants included all the songwriting candidates taking the Sorayesh practical test. The analyzed data has been obtained from the evaluation forms filled out by four evaluators for each candidate.

Results: The findings show that although the correlation coefficient among raters was high (more than 0/90), the agreement of the raters in terms of the Kappa coefficient was average. Furthermore, based on the Multi-Faceted Rasch model, the raters strictness parameter was moderate.

Conclusion: The difference between the correlation and the Kappa coefficient shows that these two indicators cannot be used alone to analyze the rater. Also, these indicators present the group status of the raters while the Multi-Faceted Rasch model shows the individual status of each rater. The results of the Multi-Faceted Rasch model indicated that the raters didn't exhibit strictness or leniency errors.

Keywords: Multi-Faceted Rasch - Rasch model - Classical Test Theory - Rater - practical test - Kappa

Cite this article: Naji, Seyedeh Hoda; Moghadamzadeh, Ali; Izanlo, Balal; Khodaie, Ebrahim (2024). Multi-Faceted Rasch Model in practical tests: Study subject: Sorayesh test. *Educational Measurement and Evaluation Studies*, 14 (47): 07-26 pages. DOI:10.22034/emes.2024.2003752.2479

Publisher: National Organization of Educational Testing (NOET)

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

مدل راش چند وجهی در آزمون‌های عملی: مورد مطالعه آزمون سُرایش

سیده هدی ناجی[۱]، علی مقدم‌زاده[۲]، بلال ایزانلو[۳]، ابراهیم خدایی[۴]

۱. دانشجوی دکتری، گروه آموزشی روش‌ها و برنامه‌های آموزشی، دانشکده روانشناسی و علوم تربیتی، دانشگاه تهران، ایران. رایانامه: s.hodanaji@ut.ac.ir
۲. گروه آموزشی روش‌ها و برنامه‌های آموزشی، دانشکده روانشناسی و علوم تربیتی، دانشگاه تهران، ایران؛ (نویسنده مسئول)، رایانامه: amoghadamzadeh@ut.ac.ir
۳. گروه آموزشی برنامه‌ریزی درسی، دانشکده روانشناسی و علوم تربیتی، دانشگاه خوارزمی، کرج، ایران. رایانامه: izan.b@khu.ac.ir
۴. گروه آموزشی روش‌ها و برنامه‌های آموزشی، دانشکده روانشناسی و علوم تربیتی، دانشگاه تهران، تهران، ایران. رایانامه: khodaie@ut.ac.ir

اطلاعات مقاله

نوع مقاله:
مقاله پژوهشی

دریافت:۱۴۰۳/۰۱/۰۵
اصلاح: ۱۴۰۳/۰۵/۲۰
پذیرش: ۱۴۰۳/۰۶/۱۵
انتشار :۱۴۰۳/۰۷/۱۵

چکیده

هدف: هدف این پژوهش تحلیل داده‌های حاصل از آزمون‌های عملی سازمان سنجش با استفاده از مدل راش چندوجهی و مقایسه آن با نتایج حاصل از تحلیل‌های کلاسیک است.

روش پژوهش: روش پژوهش کمی و از نوع توصیفی-تحلیلی است. مشارکت کنندگان شامل تمامی داوطلبان رشته آهنگسازی حاضر در آزمون عملی سُرایش بودند. داده مورد تحلیل از فرم‌های ارزیابی پر شده توسط چهار ارزیاب برای هر داوطلب بدست آمده است.

یافته‌ها: یافته نشان می‌دهد اگر چه ضریب همبستگی میان ارزیاب‌ها بالا بوده (بیش از ۰/۹۰)، توافق ارزیاب‌ها براساس ضریب کاپا، در حد متوسط است. همچنین براساس نتایج راش چند وجهی، پارامتر سخت‌گیری ارزیاب‌ها در حد متوسط بوده است.

نتیجه‌گیری: تمایز بین ضریب همبستگی و کاپا، نشان می‌دهد که نمی‌توان از هر یک از این شاخص‌ها به تنهایی برای تحلیل ارزیاب استفاده کرد، همچنین این ضرایب، وضعیت گروهی ارزیاب‌ها را نشان می‌دهند در حالی که راش چند وجهی وضعیت هر یک از ارزیاب‌ها را نشان می‌دهد. نتایج راش چند وجهی نشان داد که ارزیاب‌ها دچار خطای سخت‌گیری و سهل‌گیری نیستند.

واژه‌های کلیدی: راش چند وجهی ـ مدل راش ـ نظریه کلاسیک اندازه‌گیری ـ ارزیاب ـ آزمون عملی ـ کاپا

استناد: ناجی، سیده هدی؛ مقدم‌زاده، علی؛ ایزانلو، بلال؛ خدایی، ابراهیم (۱۴۰۳). مدل راش چند وجهی در آزمون‌های عملی: مورد مطالعه آزمون سُرایش. *مطالعات اندازه‌گیری و ارزشیابی آموزشی*، ۱۴ (شماره۴۷)، ۲۶-۷ صفحه. DOI:10.22034/emes.2024.2003752.2479
ناشر: سازمان سنجش آموزش کشور

مقدمه

در بسیاری از جوامع، سنجش و اندازه‌گیری آموزشی و روانشناختی، سهم بسیار زیادی در علوم شناختی و رفتاری دارد و منابع اطلاعاتی اساسی و قابل توجهی در مورد افراد و گروه‌ها فراهم می‌آورد. استفاده درست از آزمون‌ها می‌تواند زمینه تصمیم‌گیری بهتری در مورد افراد و برنامه‌ها را فراهم کند و راه دسترسی عادلانه‌تری به آموزش در سطوح بالاتر ایجاد کند (انجمن تحقیقات آموزشی آمریکا[1] و همکاران، ۱۳۹۸/۲۰۱۴).

در زمینه آموزش و روانشناسی، نگرانی در مورد قضاوت ارزیاب‌ها رو به افزایش است. اثرات ارزیاب مانند شدت یا ملایمت و گرایش مرکزی، معمولاً به عنوان منبع واریانس مرتبط با طرح ارزیابی در نظر گرفته می‌شود، یعنی به عنوان منبعی از واریانس نظامدار در رتبه‌بندی‌های مشاهده شده که به ارزیاب‌ها مرتبط است نه با آزمون شوندگان (لی و چا[2]، ۲۰۱۶). آزمون‌های عملکردی، در بسیاری از موارد مستلزم استفاده از چند ارزیاب برای ارزیابی عملکرد آزمودنی‌ها هستند تا قابلیت اطمینان و عینی بودن رتبه‌بندی‌ها را افزایش دهند و خطاهای ارزیاب را به حداقل برسانند، که همین مساله سبب پیچیدگی فرایند ارزیابی می‌شود (هُمبو[3] و همکاران، ۲۰۰۱؛ روبیتز و استینفلد[4]، ۲۰۱۸). آزمون شوندگان به سوالات پاسخ می‌دهند و ارزیاب‌ها طبق درک خود از سازه مورد اندازه‌گیری، ابزار اندازه‌گیری و اصول و قواعد ارزیابی مورد استفاده، پاسخ آزمون شوندگان را ارزیابی می‌کنند (اِکس[5]، ۲۰۱۵). بنابراین می‌توان گفت فرایند ارزیابی آزمون شوندگان با استفاده از ارزیاب انسانی، فرآیندی پیچیده و غیر مستقیم است. با توجه به اینکه آزمون شوندگان به طور طبیعی از نظر توانایی مورد اندازه‌گیری متفاوت هستند، این انتظار وجود ندارد که همه آن‌ها امتیاز یکسانی دریافت کنند؛ بلکه انتظار می‌رود در رتبه‌بندی توانایی‌های آزمون شوندگان اختلاف وجود داشته باشد. هر گونه تغییر در رتبه‌های آزمون شوندگان که به دلیل تفاوت قابل اعتماد در توانایی‌های آن‌ها باشد، مطلوب است. با این حال، رتبه‌بندی‌ها تحت تأثیر چندین عامل خارجی مانند نوع سوالات، رتبه‌دهندگان، شرایط و موارد دیگری قرار خواهند گرفت (لی و چا، ۲۰۱۶). اگر هنگام ارزیابی عملکرد توسط ارزیاب یک یا چند سوگیری رخ دهد، میزان سوگیری در پیش‌بینی عملکرد افراد زیاد خواهد بود در نتیجه این پیش‌بینی‌ها باعث اندازه‌گیری‌های غیرقابل اعتماد می‌شوند. سوگیری ارزیاب به واریانس نامربوط سازه نسبت داده می‌شود که مستقیما اعتبار اندازه‌گیری را تهدید می‌کند. بنابراین، چگونگی معرفی واریانس نامربوط سازه توسط ارزیاب‌ها مورد بحث است (یشیل چِنَر و ساتا، ۲۰۲۱).

بطور خلاصه اندازه‌گیری پیشرفت تحصیلی، سنجش استعداد تحصیلی و شناخت میزان توانایی علمی و مهارتی افراد، از جمله مهمترین موقعیت‌های تحقیقاتی و بررسی‌های آموزشی است. در بسیاری از اندازه‌گیری‌ها و موقعیت‌های اندازه‌گیری، ناگزیر به استفاده از آزمون‌های تشریحی و عملی خواهیم بود. استفاده از این آزمون‌ها، عموما نیازمند استفاده از ارزیاب‌های انسانی و نمره‌گذاری‌های ذهنی هستند. در این گونه نمره‌گذاری‌ها همواره اثر ارزیاب بر روی برآورد توانایی افراد وجود دارد. هر چند با آموزش ارزیاب‌ها و تهیه دستورالعمل‌های ارزیابی تلاش می‌شود اینگونه نمره‌گذاری‌ها به نمره‌گذاری عینی نزدیک شود، اما پژوهش‌ها نشان داده که بدلیل ماهیت انسان بودن ارزیاب‌ها و در نتیجه وجود خطاهای انسانی، هرگز نمی‌توان چنین اندازه‌گیری‌هایی را دقیقا عینی در نظر گرفت، بنابراین نظریه‌های مختلف اندازه‌گیری به بحث اثر ارزیاب و یا خطای ارزیاب پرداخته‌اند و تلاش کرده‌اند در نمره‌گذاری‌هایی که با واسطه ارزیاب انسانی انجام می‌شود، دقت و اعتبار نمرات و توانایی برآورد شده را افزایش دهند.

در ایران برای سنجش داوطلبان ورود به آموزش عالی در برخی از رشته‌های مهارتی مانند موسیقی، علاوه‌بر کنکور از آزمون‌های عملی یا تشریحی استفاده می‌شود. با توجه به اینکه در این آزمون‌ها از ارزیاب برای ارزیابی داوطلبان و یا تصحیح اوراق استفاده می‌شود، بررسی روانسنجی چنین آزمون‌هایی با استفاده نظریه‌های جدید اهمیت بسیاری دارد، بنابراین در این پژوهش ویژگی‌های سوال، توانایی افراد و وضعیت ارزیابان با استفاده از نظریه کلاسیک و راش چندوجهی تحلیل شده و مورد مقایسه قرار گرفته است.

1. American Educational Research Association
2. Lee and Cha
3. Hombo
4. Robitzsch and Steinfeld
5. Eckes

مبانی نظری و پیشینه پژوهش

در نظریه کلاسیک آزمون (CTT[1])، نمره مشاهده شدهٔ آزمون (X) مساوی مجموع دو جزء است: نمره واقعی (T) و نمره خطای تصادفی (E).

$$X = T + E$$

بر مبنای این نظریه نمرهٔ مشاهده شده قابل رؤیت و نمرات واقعی و خطا دارای ساخت نظری و غیر قابل مشاهده هستند. برای افزایش قابلیت اطمینان نتایج حاصل از آزمون، باید واریانس ناشی از خطای اندازه‌گیری (واریانس نمره خطا) را کاهش داد تا بتوان واریانس نمرات مشاهده شده را به واریانس نمرات واقعی نسبت داد. برای انجام این کار، معمولاً اندازه‌گیری‌ها در بعضی از شرایط از پیش تعیین شدهٔ اندازه‌گیری انجام می‌گیرد و میانگین اندازه‌گیری‌های انجام شده به عنوان یک برآورد از اندازه‌گیری "ایده‌آل" در نظر گرفته می‌شود. باید توجه داشت، تمام اندازه‌گیری‌های انجام شده برای دستیابی به نمرهٔ واقعی، باید در شرایط یکسان اندازه‌گیری باشند علاوه بر این، نتایج حاصل را تنها به همان شرایط می‌توان تعمیم داد. از آنجا که برقرار کردن این شرایط یکسان همواره امکان پذیر نیست، استفاده از نتایج اندازه‌گیری محدود می‌شود. از سوی دیگر برای تعیین اعتبار[2] اندازه‌گیری، روش‌های مختلفی در نظریه کلاسیک وجود دارد از جمله همبستگی میان نمرات دو آزمون موازی و این در حالی است که ایجاد دو آزمون موازی دارای ابهامات و دشواری‌های اساسی است، همچنین روش‌های مختلف برآورد پایایی آزمون، منجر به ضرایب پایایی مختلف و در نتیجه خطای اندازه‌گیری مختلف برای یک آزمون می‌شود و نظریه کلاسیک در تعیین مناسبترین ضریب پایایی، ناتوان است. علاوه‌براین، نظریه کلاسیک، هنگامی که طرح اندازه‌گیری پیچیده می‌شود (برای مثال آزمون‌های عملکردی و آزمون‌های غیر عینی)، نمی‌تواند پاسخگو باشد (تیلور[3]، ۱۳۹۸/۲۰۱۳). در چنین موقعیت‌هایی، بدلیل اینکه منابع اثرگذار بر واریانس ناخواسته مانند ارزیاب به طرح اندازه‌گیری اضافه می‌شود، احتمال خطای اندازه‌گیری افزایش پیدا می‌کند بنابراین باید با استفاده از روش‌هایی اعتبار اندازه‌گیری انجام شده را نشان داد. در نظریه کلاسیک اندازه‌گیری، در چنین موقعیت‌هایی، علاوه بر تحلیل سوال و بررسی پایایی ابزار، ارزیاب نیز مورد بررسی قرار می‌گیرد. برای بررسی ارزیاب در نظریه کلاسیک اندازه‌گیری، روش‌های مختلفی برای بررسی اعتبار ارزیابی‌ها وجود دارد. ضریب توافق، ضریب کاپا[4]، ضریب همبستگی پیرسون، و مقایسه میانگین ارزیاب‌ها براساس آزمون t یا تحلیل واریانس یک راهه[5]، از جمله روش‌های مورد استفاده در نظریه کلاسیک اندازه‌گیری است (ووگان[6]، ۱۹۹۱ نقل در پُلِت[7] و همکاران، ۲۰۲۲). همانگونه که توماس اِکس بیان کرده است، یک مشکل اساسی در رویکردهای سنتی وجود دارد که می‌توان نام آن را پارادوکس توافق–دقت[8] گذاشت. اِکس در پژوهش خود نشان می‌دهد که پایایی بالا و همچنین توافق بالا در بین ارزیابان، لزوما به معنای دقت بالا در ارزیابی مهارت آزمون شوندگان نیست، بنابراین پایایی و یا ضریب توافق بالا در بین ارزیابان می‌تواند منجر به نتیجه‌گیری‌های اشتباه شود. از سوی دیگر پایین بودن توافق در بین ارزیابان موجب نتیجه‌گیری نادرست در مورد ارزیابی‌ها می‌شود و همان‌گونه که در بسیاری از موقعیت‌های ارزیابی انجام می‌شود، این گونه ارزیابی به عنوان ارزیابی‌های نامناسب شناسایی شده و ارزیاب‌های ناهمسو با گروه ارزیابی، با وجود داشتن دقت بالا در ارزیابی، بدلیل همسو نبودن با سایرین، از گروه ارزیابی حذف و افراد دیگری جایگزین می‌شوند تا ضریب توافق افزایش یابد (اِکس، ۲۰۰۹). اگر چه ممکن است ارزیاب‌ها با یکدیگر موافق باشند، ولی به طور دسته جمعی برداشت اشتباهی از دستورالعمل نمره‌گذاری داشته باشند. این مساله تناقض در رفتار ارزیاب تلقی می‌شود. حتی وقتی به نظر می‌رسد داده‌های مربوط به توافق بین ارزیابان قدرت بالایی دارد، اگر عدم توافق بین ارزیابان به هر نحو با سوگیری همراه باشد، نمرات متناقض می‌تواند بسته به ارزیاب، منجر به نمره کل متفاوتی برای آزمودنی‌ها شود (تیلور، ۱۳۹۸/۲۰۱۳)، بنابراین در نظریه کلاسیک نمی‌توان عملکرد افراد را مستقل از عملکرد ارزیاب محاسبه کرد و همان‌گونه که نمرهٔ افراد وابسته به ویژگی‌های روانسنجی سوال و ویژگی‌های سوال وابسته به گروه آزمودنی است، در مورد ارزیاب نیز این وابستگی وجود دارد. اِکس راه حل این تناقض را استفاده از رویکردهای جدید اندازه‌گیری بجای رویکردهای کلاسیک

1. Classical Test Theory
2. reliability
3. Taylor
4. Kappa
5. Analysis of Variance
6. Vaughan
7. Polat
8. agreement-accuracy

می‌داند (اِکس، ۲۰۰۹). یکی از رویکردهای جدید در تحلیل ارزیاب‌ها، استفاده از مدل‌های اندازه‌گیری راش[۱] (RM) است. مدل‌های اندازه‌گیری راش از جمله مدل‌های مبتنی بر نظریه صفت مکنون[۲]، است که در سال ۱۹۶۰ توسط روان‌سنج دانمارکی، جُرج راش[۳]، معرفی گردید. راش مدل و رویکرد خود در اندازه‌گیری‌های اجتماعی را به عنوان مدل‌های احتمالی برای برخی از اندازه‌گیری‌های هوش و پیشرفت، مطرح کرد (وبند و هوا[۴]، ۲۰۲۲). این رویکرد در واقع خانواده‌ای از مدل‌های آماری است که بجای محاسبهٔ نمره در نظریهٔ کلاسیک، احتمال پاسخ درست افراد را با توجه به میزان توانایی فرد و دشواری سوال برآورد می‌کند. فرمول عمومی مدل راش بصورت زیر است:

$$P(\theta) = (\frac{e^{(\theta_n - \delta_i)}}{۱ + e^{(\theta_n - \delta_i)}})$$

که در آن:

$P_\theta =$ احتمال پاسخ درست فرد n به سوال i؛

$P_{\theta_{-۱}} =$ احتمال پاسخ نادرست فرد n به سوال i؛

$\theta_n =$ پارامتر توانایی فرد n؛

$\delta_i =$ پارامتر دشواری سوال i؛ است.

فرمول ارائه شده مدل سادهٔ راش است و برای تحلیل سوالاتی استفاده می‌شود که نمره‌گذاری آن‌ها بصورت صفر و یکی و یا به عبارت دیگر دو ارزشی[۵] است. در کنار مدل‌های دو ارزشی، مدل‌های چند ارزشی[۶] برای تحلیل داده‌های چند ارزشی ایجاد شده‌اند. مدل‌های خانوادهٔ راش می‌توانند براساس الگوی پاسخ افراد، با استفاده از روش‌های آماری و ریاضی مانند برآورد بیشینه درستنمایی[۷] (ML)، بیشینه پسین[۸] (MAP) و یا پسین مورد انتظار[۹] (EAP)، پارامتر سوال و فرد را مستقل از یکدیگر برآورد کند (امبرستون و رایس[۱۰]، ۱۳۸۸/۲۰۰۰).

خلاصه‌ای از مدل‌های چند ارزشی در جدول یک ارائه شده است (دی‌آیالا[۱۱]، ۲۰۲۲، آندریچ و ماریس[۱۲]، ۲۰۱۹). در مدل‌های راش چند ارزشی، پارامتر شیب برای سوالات درنظر گرفته نمی‌شود و احتمال پاسخ براساس پارامتر توانایی و پارامتر دشواری برآورد می‌شود، اما در مدل‌های غیر راش[۱۳]، پارامتر شیب برای سوالات در نظر گرفته می‌شود.

1. Rasch Measurement
2. Latent Theory
3. Georg Rasch
4. Wind and Hua
5. dicotomous
6. polytomous
7. maximum likelihood
8. maximum aposteriori
9. expected aposteriori
10. Embreston and Reise
11. de Ayala
12. Andrich and Marais
13. Non-Rasch

جدول ۱. مدل‌های چندارزشی

مدل‌های راش	مدل اعتبار جزئی[1] (PCM)	برآورد پارامتر توانایی – برآورد پارامتر دشواری مورد استفاده برای داده‌های گسستهٔ طبقه‌ای (نیازی به یکسان بودن طبقات پاسخ در بخش‌های مختلف آزمون ندارد) مناسب برای پاسخ‌های چند مرحله‌ای
	مدل پاسخ مرتب شده[2] (RSM)	برآورد پارامتر توانایی – برآورد پارامتر دشواری مورد استفاده برای داده‌های گسستهٔ طبقه‌ای (نیازی به یکسان بودن طبقات پاسخ در بخش‌های مختلف آزمون دارد) مناسب برای پاسخ‌های طیفی مانند طیف لیکرت
مدل‌های غیر راش	مدل اعتبار جزئی تعمیم‌یافته[3] (GPCM)	برآورد پارامتر توانایی – برآورد پارامتر دشواری – برآورد پارامتر تشخیص مورد استفاده برای داده‌های گسستهٔ طبقه‌ای (نیازی به یکسان بودن طبقات پاسخ در بخش‌های مختلف آزمون ندارد) مناسب برای پاسخ‌های چند مرحله‌ای
	مدل پاسخ مدرج[4] (GRM)	برآورد پارامتر توانایی – برآورد پارامتر دشواری – برآورد پارامتر تشخیص مورد استفاده برای داده‌های گسستهٔ طبقه‌ای (نیازی به یکسان بودن طبقات پاسخ در بخش‌های مختلف آزمون ندارد) مناسب برای پاسخ‌های چند مرحله‌ای و طیفی مانند لیکرت
	مدل پاسخ اسمی[5] (NR)	برآورد پارامتر توانایی – برآورد پارامتر دشواری – برآورد پارامتر تشخیص مورد استفاده برای داده‌های گسستهٔ که دسته‌های پاسخ در آن ذاتاً مرتب نیست مناسب برای پاسخ‌های اسمی مانند نگرش‌سنج‌ها

علاوه بر مدل‌های چندارزشی، مدل‌های پیچیده‌تر راش نیز برای تحلیل داده‌های حاصل از اندازه‌گیری‌های دارای ارزیاب ایجاد شدند. بطور کلی در هر اندازه‌گیری، علاوه بر ویژگی‌های فرد و سوال، وجوه دیگری مانند ارزیاب، جنسیت، نوع سوالات، زمان آزمون و ...، می‌تواند بر اندازه‌گیری صورت گرفته اثر بگذارد که با استفاده از مدل‌های راش چند وجهی[6] (MFRM)، می‌توان وجوه مختلف شناسایی شده را تحلیل کرد. بطور خاص تجزیه و تحلیل MFRM نشان می‌دهد که چگونه می‌توان سخت‌گیری یا سهل‌گیری رتبه‌دهنده را برآورد، درجه ثبات رتبه‌دهنده را ارزیابی، نمره هر آزمودنی تحت تاثیر سخت‌گیری رتبه‌دهنده را اصلاح و عملکرد مقیاس رتبه‌بندی را بررسی کرد، همچنین می‌توان تعاملات بالقوه میان وجوه مختلف را تشخیص داد (اِکس، ۲۰۰۹، وینِد و هوا، ۲۰۲۲).

در راش چند وجهی، برای محاسبه احتمال پاسخ صحیح، علاوه بر ویژگی آزمودنی (پارامتر توانایی) و ویژگی‌های سوال (پارامتر سوال)، پارامترهای مرتبط با رتبه‌دهنده (پارامتر سخت‌گیری) نیز در نظر گرفته می‌شود (دی‌آیالا، ۲۰۲۲، لیناکر، وینِد و هوا، ۲۰۲۲). فرمول عمومی راش چند وجهی بصورت زیر است:

1. Partial Ceredit Model
2. Rating Response Model
3. Generaliz Partial Ceredit Model
4. Graded Response Model
5. Nominal Response Model
6. Multi-Faceted Rasch Model

$$ln\left[{P_{nijk}}/{P_{nijk-۱}}\right] = \theta_n - \beta_i - \alpha_j - \tau_k$$

که در آن:

P_{nijk} = احتمال دریافت رتبه k در سوال i توسط ارزیاب j؛

$P_{nijk-۱}$ = احتمال دریافت رتبه k-۱ در سوال i توسط ارزیاب j؛

θ_n = پارامتر توانایی فرد n؛

β_i = پارامتر دشواری سوال i؛

α_j = پارامتر سخت‌گیری ارزیاب j؛

τ_k = دشواری دریافت رتبه k نسبت به دریافت رتبه k-۱؛ است.

مدل‌های مختلف راش را می‌توان به مدل چند وجهی گسترش داد. با توجه به اینکه در این مدل‌ها پارامتر شیب برای همه سوالات یکسان در نظر گرفته می‌شود، در مدل‌های چندوجهی آن‌ها نیز پارامتر تشخیص برای همه سوالات یکسان است.

پُلت و همکاران[1] (۲۰۲۲) در پژوهش خود با عنوان مقایسه نظریه کلاسیک اندازه‌گیری و نظریه راش چند وجهی، به مقایسه نظریه راش چند وجهی با نظریه کلاسیک اندازه گیری پرداخته است. در این مقاله پلت علاوه بر معرفی روش‌های کلاسیک تحلیل اثر ارزیاب، به بررسی نتایج تحلیل کلاسیک و راش چند وجهی بر روی داده‌های واقعی آزمون مهارت نوشتاری پرداخته است. پلت برای بررسی کلاسیک ارزیاب‌ها از همبستگی و آزمون F برای بررسی همسانی درونی و معنی‌داری اختلاف رتبه‌بندی ارزیاب‌ها استفاده کرده است. نتایج تحلیل کلاسیک نشان داد علارغم وجود همبستگی بالا (بیشتر از ۰/۷و یا نزدیک به آن)، اختلاف معنی‌داری بین رتبه‌بندی ارزیاب‌ها وجود دارد. همچنین نتایج راش چند وجهی نشان داد پارامتر شدت ارزیاب از ۰/۴۶- تا ۰/۴۴ متغیر بوده است. پژوهش ایزانلو و حاجت‌پور (۱۴۰۲) با عنوان بررسی رتبه‌دهی ارزیابان آزمون‌های عملکردی سراسری (طراحی صنعتی، شناخت موسیقی، نمایش عروسکی، طراحی معماری و اسکیس معماری) بر اساس روش‌های کلاسیک و مدل های چندوجهی راش، نشان داد که علارغم وجود همبستگی قابل قبول بین ارزیاب‌ها، میزان توافق بین آن‌ها قابل قبول نیست.

در پژوهش حاضر علاوه بر بررسی وضعیت ارزیابان براساس نظریه کلاسیک و راش چندوجهی، ویژگی روانسنجی سوالات نیز براساس هر دو نظریه مورد بررسی قرار گرفته است.

روش پژوهش

سازمان سنجش آموزش کشور در فرآیند سنجش داوطلبان مقطع کارشناسی ارشد رشته‌های موسیقی، علاوه بر آزمون تستی از آزمون عملی نیز استفاده می‌کند. با توجه به مطالبی که پیش از این بیان شد و با توجه به استفاده از ارزیاب در آزمون‌های سازمان سنجش، اعتبار نتایج این آزمون‌ها همواره در معرض خطای ارزیاب وجود دارد. با توجه به اهمیت نتایج آزمون‌های عملی سازمان سنجش در پذیرش و یا عدم پذیرش داوطلبان، بررسی و بهبود اعتبار نتایج این آزمون‌ها و نیز تعیین وضعیت ارزیاب‌ها با هدف بهبود ارزیابی‌ها، بسیار ضروری است. بر همین اساس در این پژوهش داده‌های سازمان سنجش آموزش کشور در آزمون عملی سُرایش برای داوطلبان رشتهٔ آهنگسازی، مورد تحلیل قرار گرفته است.

1. Polat

تحلیل‌های انجام شده با توجه به هدف توصیفی – تحلیلی بوده و از جهت کاربرد، پژوهش کاربردی محسوب می‌شود. داده‌ها براساس نظریه کلاسیک اندازه‌گیری و مدل راش چند وجهی مبتنی بر مدل اعتبار جزئی (PCMFRM[1]) و با استفاده از نرم‌افزارهای R نسخه ۴,۰/۳ بسته TAM نسخه ۴-۴/۱، SPSS نسخه ۲۷ و اِکسل نسخه ۲۰۱۶، تحلیل شده‌اند.

آزمون عملی سُرایش در سه رشتهٔ موسیقی جهانی، موسیقی ایرانی و آهنگسازی وجود دارد. این آزمون شامل سه بخش خواندن ملودی در گام‌های ماژور، خواندن ملودی مینور و وزن خوانی ترکیبی و ساده است که در رشته آهنگسازی هر بخش ۵ نمره دارد. در این آزمون هر داوطلب توسط ۴ ارزیاب مورد ارزیابی قرار می‌گیرد و هر ارزیاب نمرات خود را در فرم جداگانه ثبت می‌نماید. با توجه به اینکه تمام داوطلبان توسط هر چهار ارزیاب مورد ارزیابی قرار می‌گیرند و هر ارزیاب داوطلب را در هر سه بخش ارزیابی می‌کند، طرح ارزیابی مورد استفاده، طرح ضربدری[2] (کامل) است.

با توجه به اینکه مدل اعتبار جزئی، فقط امکان تحلیل داده‌های چند ارزشی و گسسته را دارد و از سوی دیگر، با توجه به پیوسته بودن نمرات آزمون سُرایش، لازم است پیش از اجرای تحلیل، داده‌ها از حالت پیوسته به گسسته تبدیل شوند، که با توجه به فراوانی نمرات داده شده توسط ارزیاب‌ها، نمره صفر در طبقه ۱ و سایر نمرات به ۱۰ طبقه با فاصله نیم نمره تبدیل شده‌اند (پیوست ۱).

یافته‌ها

یافته‌های پژوهش در سه بخش ارائه شده است و در هر بخش ابتدا نتایج تحلیل کلاسیک و سپس نتایج راش چندوجهی گزارش شده است.

تحلیل بخش‌های آزمون

ضرایب دشواری و تشخیص بخش‌های آزمون سُرایش در جدول شمارهٔ دو آمده است. براساس نتایج حاصل، دشواری بخش‌های مختلف آزمون از ۰/۳۸ تا ۰/۶۰ متغیر بوده است. در نظریهٔ کلاسیک اندازه‌گیری، ضریب دشواری سوال در دامنهٔ ۰ تا ۱ قرار دارد و به هر میزان که مقدار آن به یک نزدیک‌تر شود، یعنی سوال آسانتر است. در جدول شمارهٔ سه، طبقه‌بندی ضرایب دشواری و تشخیص سوالات ارائه شده است. براساس طبقه‌بندی ارائه شده برای ضریب دشواری، می‌توان گفت ضریب دشواری هر سه بخش آزمون سُرایش در سطح متوسطی قرار دارد.

جدول ۲. ضرایب دشواری و تشخیص بخش‌های آزمون سُرایش براساس CTT و PCMFRM

	CTT		PCMFRM	
بخش	ضریب دشواری	ضریب تشخیص	پارامتر دشواری	خطای استاندارد برآورد
خواندن ملودی در گام‌های ماژور	۰/۳۸	۰/٤۸	۰/۵۳	۰/۰٤
خواندن ملودی مینور	۰/٤۱	۰/٦۱	۰/۰٦	۰/۰۳
وزن خوانی ترکیبی و ساده	۰/٦۰	۰/۳۹	-۰/۵۹	۰/۰۵
میانگین	۰/٤٦	۰/٤۹	۰	۰/۰٤

1. Partial Ceredit Multi-Faceted Rasch Model
۲. Crossed Desinge

جدول ۳. طبقه‌بندی دشواری و تشخیص براساس CTT و PCMFRM

PCMFRM		CTT			
پارامتر دشواری		ضریب تشخیص		ضریب دشواری	
خیلی آسان	b ≤ -۲	ضعیف	۰ < D ≤ ۰/۰۲	بسیار سخت	P ≤۰/۳۰
آسان	-۲< b ≤+۰/۵۰	متوسط	۰/۲۰ < D ≤ ۰/٤۰	متوسط	۰/۳۰< P ≤ ۰/۸۰
متوسط	-۰/۵ < b ≤+۰/۵	خوب	۰/۰٤ < D	بسیار آسان	۰/۸۰< P
سخت	+۰/۵ < b ≤+۲				
خیلی سخت	+۲< b				

باتوجه به طبقه‌بندی ارائه شده در جدول شمارهٔ دو، براساس تحلیل کلاسیک، بخش‌های خواندن ملودی در گام‌های ماژور و خواندن ملودی مینور از نظر ضریب تشخیص در سطح خوب و بخش وزن خوانی ترکیبی و ساده در سطح متوسط قرار دارد. سخت‌ترین بخش آزمون، بخش خواندن ملودی در گام‌های ماژور و آسان‌ترین بخش، بخش وزن خوانی ترکیبی و ساده است. همچنین بخش خواندن ملودی مینور، بیشترین ضریب تشخیص را دارد.

همان‌گونه که پیش از این بیان شد، برای تحلیل راش چند وجهی، از مدل اعتبار جزئی استفاده شده است. در مدل اعتبار جزئی، آنچه تعیین کننده است، پارامتر دشواری است، بنابراین در تحلیل‌های مبتنی بر این مدل، برای بخش‌های آزمون تنها پارامتر دشواری ارائه می‌شود.

در نظریه راش، پارامتر دشواری در پیوستاری از ۳- تا ۳+ قرار دارد و هر چقدر از سمت منفی به سمت مثبت حرکت کنیم، میزان دشواری افزایش پیدا می‌کند. مقادیر دشواری نزدیک به ۲- به سوالاتی مربوط می‌شود که خیلی آسان هستند و مقادیر دشواری ۲+ به سوالاتی مربوط می‌شود که خیلی سخت هستند.

همان‌گونه که در جدول شمارهٔ دو آمده است، پارامتر دشواری بخش‌های آزمون سُرایش، از ۰/۵۹- تا ۰/۵۳ متغیر بوده. مطابق با جدول شمارهٔ پنج می‌توان گفت، هر سه بخش از نظر پارامتر دشواری متوسط هستند، وزن خوانی ترکیبی و ساده آسان‌ترین بخش و خواندن ملودی در گام‌های ماژور سخت‌ترین بخش در این آزمون بودند.

تحلیل ارزیاب

در نظریه کلاسیک برای تحلیل ارزیاب‌ها، از شباهت نمره‌گذاری (سازگاری) و میزان توافق (اجماع) میان ارزیاب‌ها استفاده می‌شود. در مبانی نظری، ضرایب همبستگی و تاو بی.کندال به عنوان شاخص‌های سازگاری و ضرایب کاپا و ضریب توافق به عنوان شاخص‌های اجماع در نظر گرفته می‌شوند. شاخص اجماع نشان‌دهندهٔ اجماع و توافق داوران در ترتیب افراد و شاخص سازگاری نشان‌دهنده رابطهٔ خطی بین نمرات داده‌شده توسط هر ارزیاب است (اِکس، ۲۰۰۵). در پژوهش حاضر از ضریب کاپا و ضریب همبستگی پیرسون برای تحلیل ارزیاب براساس نظریهٔ کلاسیک استفاده شده است.

طبق طبقه‌بندی ارائه شده توسط ولف[1]، مقادیر کاپای کمتر از ۰/۲۰، قابل قبول نیستند (وُلف، ۱۹۹۸). طبقه‌بندی ضریب کاپا در جدول شمارهٔ چهار و ضرایب کاپای محاسبه شده برای هر بخش از آزمون سُرایش، در جدول شمارهٔ پنج نشان داده شده است.

1. Wolf

جدول ٤. طبقه‌بندی ضریب کاپا

K ≤۰۰/۰	ضعیف
۰ < K ≤۰/۲۰	کم
۰/۲۰ < b ≤ ۰/٤۰	قابل قبول
۰/٤۰ < b ≤ ۰/٦۰	متوسط
۰/٦۰ < b ≤ ۰/۸۰	زیاد
۰/۸۰ < b	عالی

جدول ۵. ضریب کاپا در هر بخش

بخش	ضریب کاپا	خطای استاندارد	سطح معنی داری
خواندن ملودی در گام‌های ماژور	۰/۵۸	۰/۰۳	<۰/۰۰۰۱
خواندن ملودی مینور	۰/۵۷	۰/۰۲	<۰/۰۰۰۱
وزن خوانی ترکیبی و ساده	۰/٤٦	۰/۰۲	<۰/۰۰۰۱

با توجه به ضرایب کاپای محاسبه شده در هر سه بخش که بیش از ۰/۴۰ بدست آمده است، می‌توان گفت ضریب کاپا برای هر سه بخش قابل قبول و در حد متوسط بوده است. این ضریب برای بخش‌های خواندن ملودی در گام‌های ماژور و خواندن ملودی مینور، بالای ۰/۵۵ است.

از دیگر ضرایب مورد استفاده برای تحلیل ارزیاب‌ها در نظریه کلاسیک، ضریب همبستگی میان ارزیاب‌ها است. ضرایب همبستگی پیرسون میان ارزیاب‌ها به تفکیک بخش‌های آزمون، در جدول شمارهٔ شش نشان داده شده است.

جدول ٦. ضریب همبستگی ارزیاب‌ها

	خواندن ملودی در گام‌های ماژور			خواندن ملودی مینور			وزن خوانی ترکیبی و ساده			
ارزیاب اول	۱			۱			۱			
ارزیاب دوم	۰/۹٦	۱		۰/۹۷	۱		۰/۹٤	۱		
ارزیاب سوم	۰/۹۷	۰/۹۵	۱	۰/۹٦	۰/۹۷	۱	۰/۹۳	۰/۹۷	۱	
ارزیاب چهارم	۰/۹۸	۰/۹۵	۰/۹۷	۱	۰/۹٦	۰/۹٦	۰/۹۸	۰/۹٤	۰/۹٤	۱

مقدار این ضریب در هر سه بخش آزمون، برای هر چهار ارزیاب بیش از ۰/۹۰ و در سطح ۰/۰۱ معنادار بدست آمده است. نتایج ضریب همبستگی نشان می‌دهد که ارزیاب‌ها در رتبه‌بندی افراد دارای توافق بسیار بالایی هستند. به عبارت دیگر اگر هر ارزیاب بطور جداگانه افراد را از ضعیف به قوی مرتب کند، ترتیب افراد در بین هر چهار ارزیاب بالای ۹۰ درصد شباهت خواهد داشت.

در مدل راش چند وجهی، پارامتر ارزیاب مانند پارامتر دشواری و توانایی، در پیوستاری از ۳- تا ۳+ قرار دارد و هر چقدر از سمت منفی به سمت مثبت حرکت کنیم، میزان سخت‌گیری ارزیاب افزایش پیدا می‌کند. مقادیر نزدیک به ۳- به ارزیاب‌های بسیار سهل‌گیر و مقادیر ۳+ به ارزیاب‌های بسیار سخت‌گیر مربوط می‌شود.

جدول ۷. پارامتر ارزیاب‌ها براساس PCMFRM

ارزیاب	پارامتر سختگیری	خطای استاندارد برآورد
ارزیاب اول	۰/۰۷	۰/۰٤
ارزیاب دوم	-۰/۰۹	۰/۰٤
ارزیاب سوم	۰/۰۰	۰/۰٤
ارزیاب چهارم	۰/۰۲	۰/۰۷

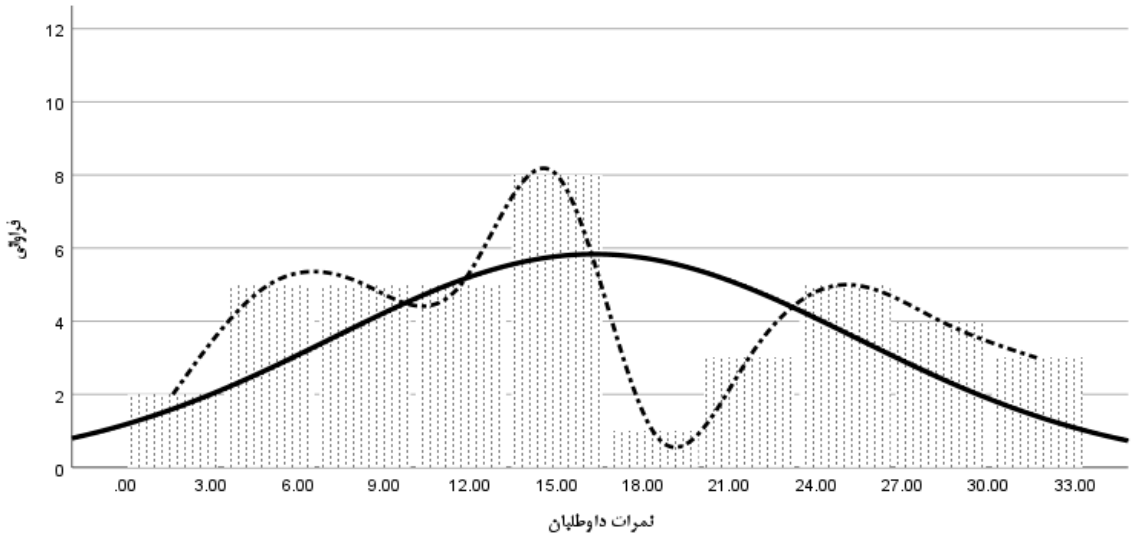
پارامتر ارزیاب برآورد شده برای هر چهار ارزیاب در جدول شمارهٔ نه آمده و نشان می‌دهد که شدت سخت‌گیری ارزیاب‌ها از ۰/۰۹- تا ۰/۰۷ متغیر بوده است. براساس نتایج بدست آمده می‌توان گفت شدت سخت‌گیری ارزیاب‌ها در حد متوسط بوده، ارزیاب اول سخت‌گیرترین و ارزیاب دوم سهل‌گیرترین ارزیاب در بین چهار ارزیاب بوده است.

تحلیل داوطلبان

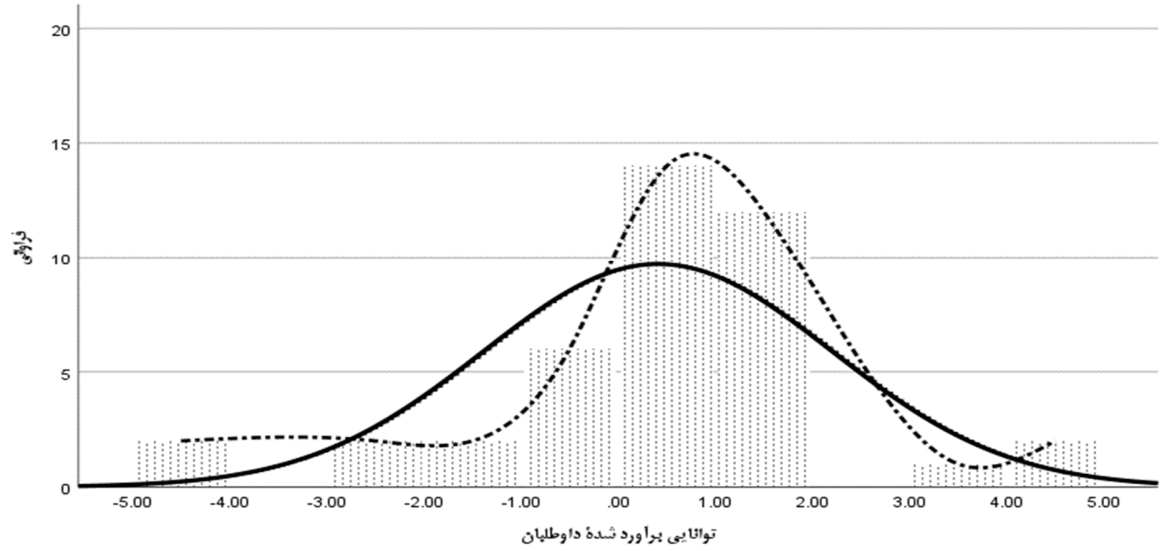
برای مقایسهٔ توانایی برآورد شده از تحلیل راش چند وجهی با نمره‌گذاری کلاسیک از ضریب همبستگی پیرسون و همچنین مقایسهٔ نمودارهای فراوانی استفاده شده است. ضریب همبستگی پیرسون ۰/۹۰ محاسبه شده است، یعنی نمرات کلاسیک می‌تواند، ۸۱ درصد از واریانس پارامتر توانایی داوطلبان را تبیین کند.

شکل‌های یک و دو، به ترتیب نمودار فراوانی نمرات کلاسیک و توانایی برآورد شده را نشان می‌دهند. همانگونه که نمودار فراوانی نمرات نشان می‌دهد، توزیع نمرات سه نمایی بوده و دارای چولگی به چپ هستند. به عبارت دیگر، بیشتر نمرات دریافت شده پائین‌تر از میانگین نمرات است. توزیع پارامتر توانایی تک نمایی بوده و به منحنی نرمال نزدیک است، با این وجود نمودار توزیع توانایی نشان می‌دهد که اکثر داوطلبان از نظر توانایی برآورد شده بالاتر از میانگین قرار داشته و از توانایی نسبتاً بالایی برخوردار هستند. نمره و پارامتر توانایی هر یک از داوطلبان در جدول شمارهٔ هشت آمده است.

شکل ۱. نمودار توزیع فراوانی نمرات براساس نظریهٔ کلاسیک



شکل ۲. نمودار توزیع فراوانی توانایی برآورد شده براساس مدل راش چند وجهی



جدول ۸ نمره و توانایی برآورد شده داوطلبان

داوطلب	توانایی	نمره	داوطلب	توانایی	نمره	داوطلب	توانایی	نمره
۱٤۰۱۱	٤/۲٤	۳۳	۱٤۰۱۱٥	۱/۰۰	۲۰/۷٥	۱٤۰۱۲۹	۰/۱۳	۱۰/۲٥
۱٤۰۱۲	٤/۲٤	۳۳	۱٤۰۱۱٦	۰/۷٤	۱۷	۱٤۰۱۳۰	-۰/۲۸	۸/۲٥
۱٤۰۱۳	۳/٥۱	۳۲/۷٥	۱٤۰۱۱۷	۰/۷۰	۱٦/٥	۱٤۰۱۳۱	-۰/۲۸	۸/۲٥
۱٤۰۱٤	۱/۹۳	۲۸/۲٥	۱٤۰۱۱۸	۰/۷۰	۱٦/٥	۱٤۰۱۳۲	-۰/۳٥	۸
۱٤۰۱٥	۱/۹۳	۲۸/۲٥	۱٤۰۱۱۹	۰/٦۸	۱٦/۲٥	۱٤۰۱۳۳	-۰/٤۲	۷/۷٥
۱٤۰۱٦	۱/۸۳	۲۷/۷٥	۱٤۰۱۲۰	۰/٦٥	۱٥/۷٥	۱٤۰۱۳٤	-۰/٥۹	۷/۲٥
۱٤۰۱۷	۱/٦۹	۲۷	۱٤۰۱۲۱	۰/٦٥	۱٥/۷٥	۱٤۰۱۳٥	-۰/۹۹	٦/۲٥
۱٤۰۱۸	۱/٤۹	۲٥/۷٥	۱٤۰۱۲۲	۰/٦۳	۱٥/٥	۱٤۰۱۳٦	-۱/٥۳	٥/۲٥
۱٤۰۱۹	۱/٤٦	۲٥/٥	۱٤۰۱۲۳	۰/٥٤	۱٤/۲٥	۱٤۰۱۳۷	-۱/۸۸	٤/۷٥
۱٤۰۱۱۰	۱/۳٦	۲٤/۷٥	۱٤۰۱۲٤	۰/٥۲	۱٤	۱٤۰۱۳۸	-۲/۳۱	٤/۲٥
۱٤۰۱۱۱	۱/۳۰	۲٤/۲٥	۱٤۰۱۲٥	۰/٤٦	۱۳/۲٥	۱٤۰۱۳۹	-۲/٥۷	٤
۱٤۰۱۱۲	۱/۲۳	۲۳/٥	۱٤۰۱۲٦	۰/٤۰	۱۲/٥	۱٤۰۱٤۰	-٤/۳٤	۳
۱٤۰۱۱۳	۱/۲۰	۲۳/۲٥	۱٤۰۱۲۷	۰/۲۰	۱۰/۷٥	۱٤۰۱٤۱	-٤/۳٤	۳
۱٤۰۱۱٤	۱/۰۷	۲۱/۷٥	۱٤۰۱۲۸	۰/۱۷	۱۰/٥			

در نظریه پرسش–پاسخ چندین روش برای مدل‌سازی اثرات ارزیاب پیشنهاد شده است که بیشتر مبتنی بر مفهوم آیتم مجازی است. آیتم مجازی مجموعه‌ای از تمام ترکیبات میان ارزیاب‌ها و آیتم‌ها است. در این پژوهش، ۳ بخش و چهار ارزیاب وجود دارد و با توجه به اینکه از طرح ارزیابی متقاطع استفاده شده است، بنابراین ۱۲ بخش مجازی (۳ بخش × ۴ ارزیاب) در تحلیل وجود دارد. در جدول شماره نه، شاخص‌های برازش مدل در دوازده بخش مجازی آمده است.

جدول ۹. شاخص برازش آیتم‌های مجازی

بخش مجازی	شماره بخش مجازی	Infit	Infit_t	Infit_p
بخش۱-ارزیاب۱	۱	۰/٦۷	-۱/۲	۰/۲۳
بخش۱-ارزیاب۲	۲	۰/٥٤	-۱/۹۷	۰/۰٥
بخش۱-ارزیاب۳	۳	۰/٤۲	-۲/٥٦	۰/۰۱
بخش۱-ارزیاب٤	٤	۰/٥٦	-۱/۷٥	۰/۰۸
بخش۲-ارزیاب۱	٥	۰/۹۰	-۰/۳۲	۰/۷٥
بخش۲-ارزیاب۲	٦	۱/۰۱	۰/۱۱	۰/۹۱
بخش۲-ارزیاب۳	۷	۱/۲۱	۰/۸۸	۰/۳۸
بخش۲-ارزیاب٤	۸	۰/٦۷	-۱/۳۸	۰/۱۷
بخش۳-ارزیاب۱	۹	۱/٤۰	۱/٥۱	۰/۱۳
بخش۳-ارزیاب۲	۱۰	۱/٤۳	۱/٦٦	۰/۱۰
بخش۳-ارزیاب۳	۱۱	۱/۸۱	۲/۸٤	۰/۰۱
بخش۳-ارزیاب٤	۱۲	۱/۲۸	۱/۱۸	۰/۲٤
میانگین		۰/۹۹		
انحراف استاندارد		۰/٤۳		

برای بررسی برازش مدل با داده‌ها از شاخص‌های برازش Infit استفاده شده است. آماره Infit تفاوت بین نمرات مشاهده شده و نمرات مورد انتظار است. چنانچه تفاوت بین نمرات مشاهده شده داوطلبان و نمرات مورد انتظار آن‌ها زیاد باشد، می‌تواند دقت اندازه‌گیری را تهدید کند. مقدار قابل قبول برای شاخص Infit در فاصله دو انحراف استاندارد از میانگین است، بنابراین دامنه قابل قبول برای این شاخص در محدوده ۱/۸۵ تا ۰/۱۳ است. براساس مقادیر Infit در جدول شماره یازده می‌توان گفت برازش مدل با داده در تمام آیتم‌های مجازی قابل قبول بوده است.

در جدول شماره ده، شاخص‌های برازش داوطلبان آمده است. مقادیر Infit که بیش از دو انحراف استاندارد با میانگین فاصله داشته باشد، نشان‌دهنده توافق کم ارزیاب‌ها در مورد داوطلب است. بنابراین در خصوص داوطلبانی که مقدار Infit آن‌ها بیش از ۲/۱۰ است، توافق کمی بین ارزیاب‌ها وجود دارد. همانگونه که در جدول مشخص است، ارزیاب‌ها در ارزیابی داوطلب شمارهٔ ۱ و ۱۰ توافق کمی داشته‌اند.

جدول ۱۰. شاخص برازش توانایی داوطلبان

داوطلب	infitPerson	داوطلب	infitPerson
۱	۲/۷٤	۲۲	۰/۱۷
۲	۱/٦٥	۲۳	۰/۳٦
۳	۰/۲٦	۲٤	۰/۰٥
٤	۰/۸۷	۲٥	۰/٥۷
٥	۱/۱۷	۲٦	۰/۳۰
٦	۱/۷٦	۲۷	۰/۲۳

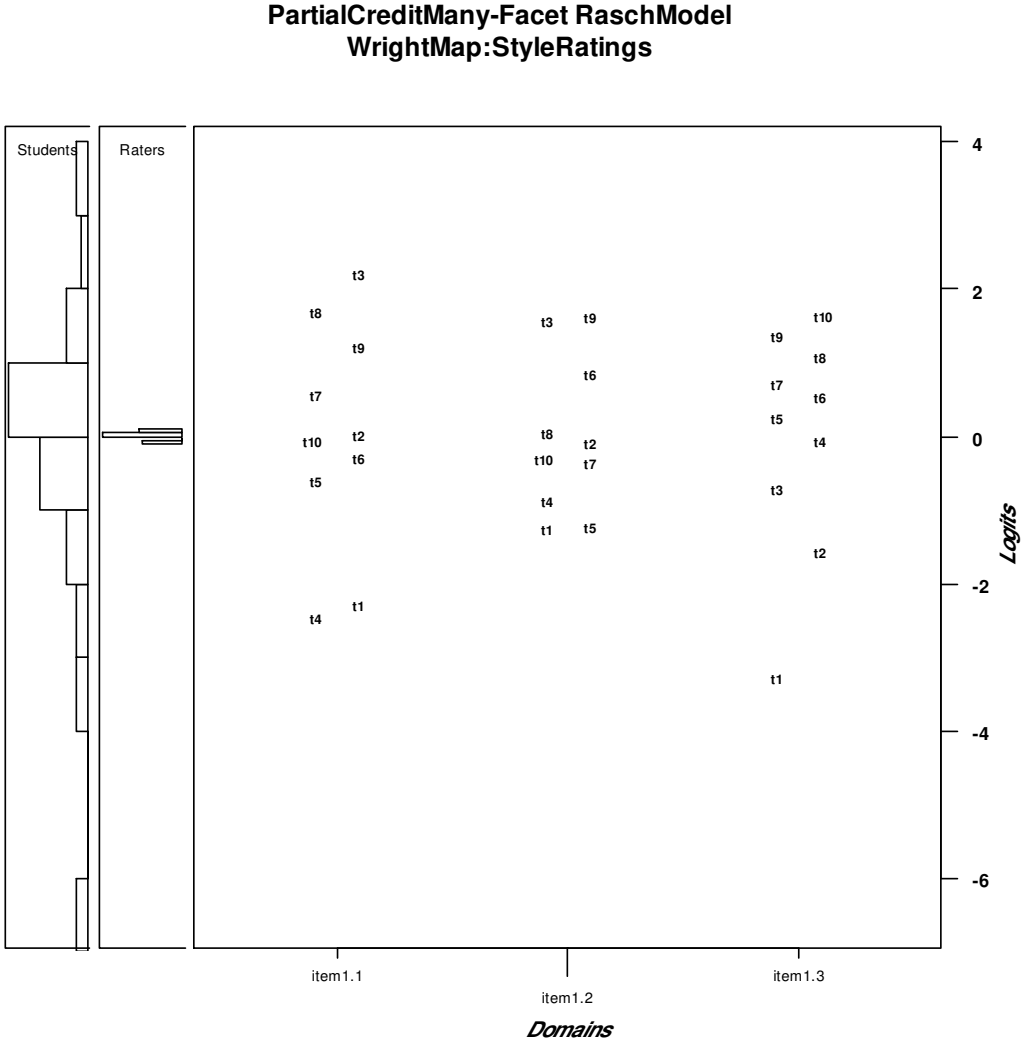
۷	۰/۹٤	۲۸	۰/۱٤
۸	۰/۰۵	۲۹	۱/۱۹
۹	۰/۵٤	۳۰	۰/۰۷
۱۰	۲/۸۷	۳۱	۱/۱۵
۱۱	۱/۱۷	۳۲	۰/٤۷
۱۲	۱/۳۰	۳۳	۰/۲٦
۱۳	۰/۷۷	۳٤	۰/۷۲
۱٤	۰/۱۷	۳۵	۰/۵۱
۱۵	۰/۷٦	۳٦	۰/٤٦
۱٦	۰/۲۵	۳۷	۰/۰۹
۱۷	۰/۹۰	۳۸	۰/۳۸
۱۸	۱/۷۹	۳۹	۰/۰۹
۱۹	۰/۷٦	٤۰	۱/۲۷
۲۰	۰/۵۱	٤۱	۰/۳۵
۲۱	۰/۲۸		
میانگین		۰/۷٤	
انحراف استاندارد		۰/٦۸	

خلاصه‌ای از وضعت پارامتر توانایی داوطلبان، پارامتر شدت ارزیاب‌ها و دشواری بخش‌های مختلف آزمون در نقشه رایت[1] قابل مشاهده است. همانگونه که در تصویر شمارهٔ دو مشاهده می‌شود، دامنهٔ توانایی داوطلبان از بسیار ضعیف تا بسیار قوی بوده و اکثر داوطلبان از نظر توانایی در سطح متوسط بودند. همچنین هر چهار ارزیاب از نظر شدت سختگیری، در حد متوسط بودند.

یکی از خروجی‌های قابل توجه در تحلیل‌های چند ارزشی راش، بررسی میزان دشواری دستیابی به هر یک از نمرات است. همانگونه که در نقشهٔ رایت مشاهده می‌شود، در بخش اول و دوم آزمون سُرایش، دشواری دریافت نمرات مختلف نامرتب است اما در بخش سوم، دشواری دریافت نمرات مرتب است به این صورت که با افزایش نمره، دشواری دریافت آن نیز افزایش یافته است.

1 Wright Map

شکل ۳. نقشهٔ رایت



بحث و نتیجه‌گیری

در این پژوهش، آزمون عملی سُرایش برای داوطلبان رشتهٔ آهنگسازی مورد تحلیل قرار گرفت. این آزمون شامل ۳ بخش ۵ نمره‌ای است و از چهار ارزیاب برای ارزیابی داوطلبان استفاده شده است. تعداد کل داوطلبان حاضر در این آزمون ۴۱ نفر بوده که فرم ارزیابی همه داوطلبان مورد تحلیل قرار گرفت. یافته‌های حاصل از تحلیل کلاسیک نشان می‌دهد که بطور کلی آزمون از لحاظ دشواری نسبتا دشوار بوده است. سخت‌ترین بخش آزمون خواندن ملودی در گام‌های ماژور و آسان‌ترین بخش، وزن خوانی ترکیبی و ساده بود. در آزمون‌های توانایی، کمترین میزان ضریب تشخیص قابل قبول در برخی منابع ۰/۲۰ و در برخی دیگر از منابع ۰/۳۰ گزارش شده است. با توجه به اینکه ضریب تشخیص هر سه بخش آزمون بالای ۰/۳۰ است، می‌توان گفت هر سه بخش از نظر ضریب تشخیص در سطح قابل قبولی قرار دارند.

برای تحلیل ارزیاب در نظریه کلاسیک اندازه‌گیری از روش‌های مختلفی همچون ضریب همبستگی، ضریب توافق، ضریب کاپا، آزمون تی یا تحلیل واریانس یک راهه، استفاده می‌شود. در این پژوهش از دو روش ضریب کاپا برای بررسی پایایی توافق ارزیابان و ضریب همبستگی برای

بررسی میزان همسویی آن‌ها در ارزیابی انجام شده، استفاده شده است. ضریب کاپای بدست آمده برای بخش خواندن ملودی در گام‌های ماژور برابر با ۰/۵۸، برای بخش خواندن ملودی مینور برابر با ۰/۵۷ و برای بخش وزن خوانی ترکیبی و ساده برابر با ۰/۴۶ بدست آمده است. مقدار ضریب کاپا عددی از ۰ تا ۱ است و مقادیر بیشتر نشان دهنده پایایی بالاتر ارزیاب‌ها است، یعنی ارزیاب‌ها در رتبه‌بندی افراد توافق بیشتری دارند. براساس طبقه‌بندی کیفی ضریب کاپا، مقادیر بین ۰/۴۱ تا ۰/۶۰ نشان‌دهنده توافق متوسط بین ارزیابان است. بر این اساس می‌توان گفت در آزمون عملی سُرایش، توافق ارزیاب‌ها در سطح متوسط است، از سوی دیگر ضریب همبستگی حاصل بین ارزیاب‌ها در هر سه بخش آزمون بیش از ۰/۹۰ است که نشان‌دهنده همبستگی بسیار بالا بین آن‌ها است. اختلاف بین ضریب کاپا و ضریب همبستگی بین ارزیاب‌ها را می‌توان چنین تحلیل کرد؛ ضریب کاپا نشان‌دهنده توافق ارزیاب‌ها در نمره اختصاص داده شده به افراد است، در حالی که ضریب همبستگی نشان‌دهنده ترتیب افراد از نظر ارزیاب‌ها است. به عبارت دیگر اگر داوطلبان را براساس ارزیابی هر یک از ارزیاب‌ها بصورت جداگانه رتبه‌بندی کنیم، رتبه‌بندی هر چهار ارزیاب یکسان خواهد بود، اما نمره‌ای که هر یک از ارزیاب‌ها داده‌اند متفاوت است، برای مثال داوطلب ۱۴۰۱۱ نمرات متفاوتی از هر ارزیاب دریافت کرده است اما از نظر هر چهار ارزیاب در رتبه اول قرار می‌گیرد. پژوهش‌های پیشین نیز اختلاف بین همبستگی و ضریب کاپا را گزارش کرده‌اند (پلت و همکاران، ۲۰۲۲ و ایزانلو و حاجت‌پور، ۱۴۰۲). واقع می‌توان گفت اگرچه ترتیب داوطلبان از نظر توانایی در بین ارزیاب‌ها یکسان است، اما الزاما نمره‌ایی که به داوطلب‌ها داده‌اند، یکسان نیست. در این پژوهش‌ها از ضریب همبستگی به عنوان توافق نسبی و از ضریب کاپا به عنوان توافق مطلق نام برده شده است. براساس نتایج تحلیل راش چند وجهی، بخش‌های مختلف آزمون سُرایش از نظر دشواری در سطح متوسطی بودند، همچنین پارامتر ارزیاب نشان می‌دهد که هر چهار ارزیاب از نظر سخت‌گیری و سهل‌گیری در میانه قرار داشتند. پارامتر ارزیاب در محدوده ۳- تا ۳+ قرار دارد که هر چه از ۳- به ۳+ نزدیکتر شود، سخت‌گیری ارزیاب بیشتر خواهد بود. با توجه به اینکه پارامتر ارزیاب برای هر چهار ارزیاب در محدود ۰/۰۹- تا ۰/۰۷ برآورد شده است، باید گفت که ارزیاب‌ها میانه‌رو بوده و دارای خطای سهل‌گیری و یا سخت‌گیری نیستند که این مسئله را می‌توان در همبستگی بالای نمرات کلاسیک و پارامتر توانایی نیز مشاهده کرد.

بطور کلی نتایج این پژوهش نشان می‌دهد که اگر چه همبستگی ارزیاب‌ها می‌تواند بالا باشد، اما ضریب همبستگی به تنهایی نمی‌تواند برای بررسی وضعیت ارزیاب‌ها شاخص مناسبی باشد، از سوی دیگر ضرایب کاپا، همبستگی و سایر شاخص‌های مبتنی بر نظریه کلاسیک برای بررسی ارزیاب، تنها می‌تواند وضعیت گروهیِ ارزیاب‌ها را از نظر همسویی و همخوانی نشان دهد، در حالی که مدل راش چند وجهی می‌تواند وضعیت هر یک از ارزیاب‌ها را از نظر میزان سخت‌گیری و یا سهل‌گیری نشان دهد. همچنین در مدل راش چند وجهی، پارامتر دشواری سوال، مستقل از عملکرد ارزیاب برآورد می‌شود، به عبارت دیگر عملکرد ارزیاب تاثیری در میزان دشواری سوال ندارد. بطور کلی اگر چه ترتیب دشواری بخش‌های آزمون در هر دو روش تحلیل یکسان است (خواندن ملودی در گام‌های ماژور، خواندن ملودی مینور، وزن خوانی ترکیبی و ساده)، اما میزان دشواری آن‌ها متفاوت بدست آمده است. براساس نتایج حاصل از نظریه کلاسیک، بخش خواندن ملودی در گام‌های ماژور سخت، خواندن ملودی مینور متوسط و وزن خوانی ترکیبی و ساده آسان محاسبه شده‌اند، در حالی که براساس نتایج راش چند وجهی، دشواری هر سه بخش در حد متوسط برآورد شده است. از سوی دیگر براساس نظریه کلاسیک، برای هر بخش ضریب تشخیص محاسبه شده است، در حالی که در تحلیل راش چند وجهی باتوجه به استفاده از مدل اعتبار جزئی، پارامتر تشخیص برای همه بخش‌ها یکسان و برابر با یک در نظر گرفته می‌شود. بنابراین می‌توان گفت در تحلیل‌هایی که میزان تشخیص بخش‌ها و یا سوالات آزمون اهمیت تعیین کننده دارد، استفاده از نظریه کلاسیک نسبت به راش چند وجهی مبتنی بر اعتبار جزئی، می‌تواند آگاهی‌دهنده‌تر باشد، اما در مورد تحلیل ارزیاب و دشواری بخش‌ها یا سوالات آزمون، تحلیل‌های راش چند وجهی آگاهی‌دهنده‌تر از نتایج کلاسیک خواهد بود. از سوی دیگر با توجه به توزیع توانایی داوطلبان، انتظار می‌رود که بیشترین فراوانی نمرات در سطح متوسط باشد، اما همان‌گونه که در نمودار یک مشخص است، بیشترین فراوانی در نمرات پائین‌تر از میانگین قرار دارد. تفاوت توزیع فراوانی نمرات با توزیع فراوانی توانایی افراد، نشان می‌دهد که تعیین سطح توانایی افراد در نظریهٔ کلاسیک تحت تاثیر وجوه مختلف از جمله دشواری سوالات و همچنین عملکرد ارزیاب قرار گیرد و افرادی که از نظر پارامتر توانایی در یک محدوده قرار دارند، ممکن است نمرات متفاوتی را دریافت کنند. اگر چه ترتیب افراد براساس نتایج کلاسیک و راش چند وجهی یکسان است، اما نمرات داده شده به داوطلبان متناسب با توانایی واقعی آن‌ها نبود که دلیل آن مستقل نبودن نمره افراد از سختی سوالات و عملکرد ارزیاب در نظریهٔ کلاسیک اندازه‌گیری است. با توجه به اینکه نظریه راش چند وجهی می‌تواند توانایی افراد را بگونه‌ای برآورد کند که تحت تاثیر عملکرد ارزیاب قرار نداشته باشد، و از سوی دیگر عدم امکان بررسی و حذف اثر ارزیاب بر روی نمره افراد در نظریه کلاسیک اندازه‌گیری، نظریه راش چند وجهی در بررسی وضعیت داوطلبان، آگاهی‌دهنده‌تر از نظریه کلاسیک اندازه‌گیری خواهد بود.

References

Allen, M. J., & Yen, W. M. (۲۰۰۲). *Introduction to Measurement Theory*. Prospect Heights, IL: Waveland Press. Translated in persian, ۲۰۱۳, SAMT

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (۲۰۱۴). *Standards for educational and psychological testing*. American Educational Research Association. Tranlated in persian, ۲۰۱۹, Tehran university.

Andrich D. & Marais I. (۲۰۱۹). *A course in rasch measurement theory: measuring in the educational social and health sciences*. Springer. https://doi.org/۱۰/۱۰۰۷/۹۷۸-۹۸۱-۱۳-۷۴۹۶-۸

de Ayala, R. J. (۲۰۲۲). *The theory and practice of item response theory* (۲st ed.). Guilford Press.

Eckes, T. (۲۰۱۵). *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments* (۲nd ed.). New York: Peter Lang.

Embretson, Susan & Reise, S.. (۲۰۰۰). *Item Response Theory For Psychologists*. Tranlated in persian, Roshd, ۲۰۰۹, Tehran.

Ezanloo, B., & Hajatpour, S. (۲۰۲۳). 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*, ۱۲(۴۲), ۱۰۰-۱۲۳. doi: ۱۰/۲۲۰۳۴/emes.۲۰۲۳/۵۲۸۱۶۱.۲۲۴۴

Hambleton, Ronald K. & Swaminathan, Hariharan. & Rogers, H. Jane. (۱۹۹۱). *Fundamentals of item response theory*. Newbury Park, Calif: Sage Publications, Translated in persian, ۲۰۱۰, Alameh tabai university.

Hombo, C. M., Donoghue, J. R., & Thayer, D. T. (۲۰۰۱). A simulation study of the effect ofrater designs on ability estimation (ETS Research Report No. RR-۰۱-۰۵). Princeton, NJ:Educational Testing Service.

Keeves J. P. (۱۹۹۷). Introduction: Advances in Measurement in Education. In Keeves J. P. (Ed.), *Educational research methodology and measurement: an international handbook* (۲nd ed.) (pp. ۷۰۵-۷۱۲). Pergamon.

Lee, M., & Cha, D. (۲۰۱۶). A comparison of generalizability theory and many facet Rasch measurement in an analysis of mathemetics creative problem solving test. *Journal of Curriculum Evaluation*, ۱۹(۲), ۲۵۱–۲۷۹

Li, G., Pan, Y., & Wang, W. (۲۰۲۱). Using Generalizability Theory and Many-Facet Rasch Model to Evaluate In-Basket Tests for Managerial Positions. *Frontiers in psychology*, ۱۲, ۶۶۰۵۵۳. https://doi.org/۱۰/۳۳۸۹/fpsyg.۲۰۲۱/۶۶۰۵۵۳

Polat, M., Sölpük Turhan, N., & Toraman, Çetin . (۲۰۲۲). Comparison of Classical Test Theory vs. F Theory in writing assessment. *Pegem Journal of Education and Instruction*, ۱۲(۲), ۲۱۳–۲۲۵. https://doi.org/۱۰/۴۷۷۵۰/pegegog.۱۲/۰۲.۲۱

Robitzsch, A., & Steinfeld, J. (۲۰۱۸). Item response models for human ratings: Overview, estimation methods, and implementation in R. *Psychological Test and Assessment Modeling*, ۶۰(۱), ۱۰۱–۱۳۹.

Taylor, C. S. (۲۰۱۳). *Validity and validation.* Oxford University Press, Translated in persian, ۲۰۱۰, Alameh tabai university.

Wind, S., & Hua, C. (۲۰۲۲). *Rasch Measurement Theory Analysis in R* (۱st ed.). Chapman and Hall/CRC. https://doi.org/۱۰/۱۲۰۱/۹۷۸۱۰۰۳۱۷۴۶۶۰

Wolf. R.M. (۱۹۹۷). Rating Scales. In Keeves J. P. (Ed.), *Educational research methodology and measurement: an international handbook* (۲nd ed.) (pp. ۹۵۸-۹۶۵). Pergamon.

پیوست ۱.

جدول ۱. فروانی رتبه‌های داده شده توسط هر ارزیاب

	خواندن ملودی در گام‌های ماژور				خواندن ملودی مینور				وزن خوانی ترکیبی و ساده			
نمره	ارزیاب ۱	ارزیاب ۲	ارزیاب ۳	ارزیاب ٤	ارزیاب ۱	ارزیاب ۲	ارزیاب ۳	ارزیاب ٤	ارزیاب ۱	ارزیاب ۲	ارزیاب ۳	ارزیاب ٤
۰	۱۰	۸	۸	۹	۱۱	۸	۱۰	۹	۳	۲	۲	۱
۰/۲۵	–	٦	۳	–	–	٤	٤	–	–	۱	۱	–
۰/۵۰	۱۰	۵	۱۰	۱۲	٤	٤	۳	۷	۲	۲	۳	۳
۰/۷۵	–	–	–	–	–	–	–	–	–	–	–	–
۱	٦	۳	۵	۳	۵	۳	٤	۲	۵	۲	٤	۱
۱/۲۵	–	–	–	–	–	–	–	–	–	–	–	–
۱/۵۰	–	–	–	۱	–	–	–	۲	۱	۵	۲	۵
۱/۷۵	–	–	–	–	–	–	–	–	–	–	–	–
۲	۲	۳	۲	۱	۱	۲	–	۱	٦	۱	۲	۹
۲/۲۵	–	–	–	–	–	–	–	–	–	–	–	–
۲/۵۰	۲	٤	۲	٤	٤	٤	۲	۳	۱	٤	٤	۵
۲/۷۵	–	–	–	–	–	–	–	–	–	–	–	–
۳	۵	٤	٤	۳	۱	۲	۲	۱	٦	۱۳	۳	۳
۳/۲۵	–	–	–	–	–	–	–	–	–	–	–	–
۳/۵۰	۱	٤	۳	۳	۳	۱	۳	٤	٤	۵	٤	٤
۳/۷۵	–	–	–	–	–	–	–	–	–	–	–	–
٤	۱	–	–	۲	٤	٤	٤	۲	۵	٤	۳	۵
٤/۲۵	–	–	–	–	–	–	–	–	–	–	–	–
٤/۵۰	–	۱	۱	–	۱	۱	۱	–	۳	۳	۵	٦
٤/۷۵	–	–	–	–	–	۱	–	–	–	۲	–	–
۵	٤	۳	۳	۳	۷	۷	۸	۳	۵	۷	۸	۳

جدول ۲. طبقه‌بندی نمرات آزمون سُرایش

طبقه	دامنه	طبقه	دامنه
۱	۰	۷	[۳، ۲/۵)
۲	[۰/۵، ۰)	۸	[۳/۵، ۳)
۳	[۱، ۰/۵)	۹	[٤، ۳/۵)
٤	[۱/۵، ۱)	۱۰	[٤/۵، ٤)
۵	[۲، ۱/۵)	۱۱	[۵، ٤/۵)
٦	[۲/۵، ۲)		