Educational Measurement
13(49), Autumn 2022, 31-51
jem.atu.ac.ir
DOI: 10.22054/jem.2023.63328.2219

ATU PRESS

Fair Classroom Assessment Rubric: Fitting a Multidimensional Graded Response Model

Ali Baniasadi — Ph.D. Student in Assessment and Measurement, University of Tehran, Tehran, Iran

Keyvan Salehi * — Assistant Professor, Curriculum Development & Instruction Methods Dept., University of Tehran, Tehran, Iran

Ebrahim Khodaie — Associate Professor, Curriculum Development & Instruction Methods Dept., University of Tehran, Tehran, Iran

Khosro Bagheri — Professor, University of Tehran, Tehran, Iran

Balal Izanloo — Associate Professor, University of Kharazmi, Tehran, Iran

Abstract

The present study aimed to investigate the psychometric properties of fair classroom assessment Rubric based on Item-Response theory. For this purpose, a sample of 511 students of the University of Tehran was selected by the available sampling method and answered Rubric questions. At this stage, to determine the application of unidimensional or multidimensional models, DETECT and parallel analysis methods were used. The results of both methods rejected the unidimensionality of the data and the results of the parallel analysis showed the extraction of three factors from the data. Also, the comparison of unidimensional and multidimensional model fit indices including log-likelihood, likelihood ratio, Root Mean Square Error of Approximation and comparison of Bayesian and Akaike information criteria confirmed the better fit of the multidimensional model for the data. Thus, due to the polytomous of the answers to the questions, the multidimensional graded response model was used to estimate the parameters of the questions. The reliability of each of the subscales of procedural fairness, nature of assessment and interactional fairness were 0.85, 0.69 and 0.63, respectively. Estimation of the discrimination parameters ranged from 1.048 to 5.802, which showed that all the questions performed well in the discrimination of the upper and lower levels of the fair classroom assessment, and after controlling the false discovery rate, the S-X2 statistic showed a good fit of all Rubric questions. In general, the results of this study show that the developed Rubric has appropriate psychometric properties to evaluate the quality of fairness in the classroom assessment.

Keywords: Rubric, Classroom Assessment, Fairness, Graded Response Model, Item Response Theory.

* Corresponding Author: keyvansalehi@ut.ac.ir

How to Cite: Baniasadi, A., Salehi, K., Khodaie, E., Bagheri, K., & Izanloo, B. (2022). Fair Classroom Assessment Rubric: Fitting a Multidimensional Graded Response Model. *Quarterly of Educational Measurement*, *13*(49), 31-51. doi: 10.22054/jem.2023.63328.2219

Original Reserch
Received: Sept. 06, 2021
Accepted: Sept. 10, 2022
ISSN: 2252-004x
eISSN: 2470-6240

مقاله پژوهشی

اندازه‌گیری تربیتی
دوره ۱۳، شماره ۴۹، پاییز ۱۴۰۱، ۳۱-۵۱
jem.atu.ac.ir
DOI: 10.22054/jem.2023.63328.2219

روبریک سنجش کلاسی عادلانه: برازش مدل چندبعدی پاسخ مدرج

علی بنی‌اسدی | دانشجوی دکتری رشته سنجش و اندازه‌گیری، دانشگاه تهران، تهران، ایران

کیوان صالحی * | استادیار، گروه روش‌ها و برنامه‌های آموزشی، دانشگاه تهران، تهران، ایران

ابراهیم خدایی | دانشیار، گروه روش‌ها و برنامه‌های آموزشی، دانشگاه تهران، تهران، ایران

خسرو باقری | استاد، گروه مبانی فلسفی و اجتماعی آموزش و پرورش، دانشگاه تهران، تهران، ایران

بلال ایزانلو | استادیار، گروه برنامه‌ریزی درسی، دانشگاه خوارزمی، تهران، ایران

تاریخ دریافت: ۱۴۰۰/۰۶/۱۵؛ تاریخ پذیرش: ۱۴۰۱/۰۶/۱۹

ISSN: 2252-004x

eISSN: 2470-6240

چکیده

هدف پژوهش حاضر بررسی ویژگی‌های روان‌سنجی روبریک سنجش کلاسی عادلانه بر اساس نظریه سؤال- پاسخ بود؛ به این منظور نمونه‌ای شامل ۵۱۱ دانشجوی دانشگاه تهران به روش نمونه‌گیری در دسترس انتخاب شدند و به سؤال‌های روبریک پاسخ دادند. در این مرحله و به‌منظور تعیین کاربرد مدل‌های تک‌بعدی یا چندبعدی، روش‌های دیتکت و تحلیل موازی به کار رفت که نتایج هر دو روش، تک‌بعدی بودن داده‌ها را رد کرده و نتایج حاصل از تحلیل موازی نشان‌دهنده استخراج سه عامل از داده‌ها بود. همچنین مقایسه شاخص‌های برازش مدل تک‌بعدی و چندبعدی شامل لگاریتم درستنمایی، نسبت درستنمایی، ریشه دوم میانگین مربعات باقی‌مانده و مقایسه ملاک‌های اطلاعات بیزی و آکایک برازش بهتر مدل چندبعدی برای داده‌ها را تائید کرد. به‌این‌ترتیب و با توجه به چندارزشی بودن پاسخ سؤال‌ها، مدل چندبعدی پاسخ مدرج برای برآورد پارامترهای سؤال‌ها به کار رفت. پایایی هر یک از زیرمقیاس‌های عدالت رویه‌ای، ماهیت سنجش و عدالت تعاملی به ترتیب ۰/۸۵، ۰/۶۹ و ۰/۶۳ به دست آمد. برآورد پارامترهای تمیز سؤال‌ها در دامنه ۱/۰۴۸ تا ۵/۸۰۲ قرار داشت که نشان داد همه سؤال‌ها در تمیز سطوح بالا و پایین سنجش کلاسی عادلانه به‌خوبی عمل می‌کنند و بعد از کنترل نرخ کشف اشتباه، آماره S-X2 مؤید برازش خوب همه سؤال‌های روبریک بود. درمجموع، نتایج این پژوهش نشان‌دهنده آن است که روبریک تدوین‌شده دارای ویژگی‌های روان‌سنجی مناسبی برای ارزشیابی کیفیت عدالت در سنجش کلاسی اساتید است.

کلیدواژه‌ها: روبریک، سنجش کلاسی، عدالت، مدل پاسخ مدرج، نظریه سؤال- پاسخ

* نویسنده مسئول: Keyvansalehi@ut.ac.ir

مقدمه

نظام آموزش عالی از بزرگ‌ترین نظام‌های درون جامعه است که سرنوشت جامعه را در بلندمدت مشخص می‌کند و دانشگاه‌ها یکی از مهم‌ترین اجزای نهاد آموزش و شکل‌دهی جهان آینده تلقی می‌شوند (کاویانی و نصر، ۱۳۹۵). اهداف و رسالت‌های مختلفی چون افزایش درک بین‌المللی، تشویق پژوهش در بالاترین سطح، پرورش مهارت‌های استدلال انتقادی و تشویق برابری اجتماعی برای این نظام که در ارتقا بخش‌های مختلف جامعه نقش مهمی ایفا می‌کند؛ تعریف شده است (ابطحی و ترابیان، ۱۳۸۹). دستیابی به این اهداف مستلزم بهبود و ارتقای کیفیت آن است که وابسته به چندین عامل ازجمله برنامه درسی است (قنبری و نیکخواه، ۱۳۹۵). ارزشیابی از آموخته‌های دانشجویان به‌عنوان ایدئولوژی عملیاتی، یکی از عناصر مهم برنامه درسی است که بر سایر عناصر آن، نحوه عملکرد مدرس، رفتارهای دانشجو و حتی ساختار نظام آموزشی اثر می‌گذارد و هدف آن شناسایی نقاط ضعف و قوت دانشجویان و ارائه بازخورد به آن‌ها، اصلاح و بازبینی شیوه‌های تدریس، اصلاح برنامه درسی، ارائه گزارش به ذینفعان و مدیران نظام آموزشی و تهیه داده‌های مورد نیاز برای تصمیم‌گیری‌ها و تعیین نمره دانشجویان مربوط می‌شود (Eisner, 1994 نقل از سراجی و همکاران، ۱۳۹۲). ETS (2009) سنجش عادلانه را کمکی به افزایش کیفیت در آموزش می‌داند و رعایت عدالت را مستلزم حساسیت نسبت به نیازها و احساسات آزمون‌دهندگان، اجتناب از تصاویر و محتوای توهین‌آمیز و نبود موانع غیرضروری برای موفقیت همه‌ی آزمون‌دهندگان می‌داند و مکانیسم دیدن این الزامات را «وارسی عدالت[1]» می‌نامد.

سنجش عادلانه ازاین‌جهت که نتایج سنجش زمینه‌های تحصیلی آینده و انتخاب‌های شغلی را تحت تأثیر قرار می‌دهد مهم است و ماهیت چنین سنجش‌هایی روی تمرین‌های کلاسی، روی انگیزش و خودتنظیمی دانش‌آموزان در یادگیری اثر داشته و می‌تواند پیامدهای اجتماعی، هیجانی و تحصیلی بلندمدتی برای دانش‌آموزان به همراه داشته باشد (DeLuca, 2012) این مکانیسم به‌گونه‌ای است که عناصر موقعیتی و ادراکی نظیر عدالت در آغاز بر اخلاق تحصیلی تأثیر گذاشته و سپس اخلاق تحصیلی موجب کاهش رفتارهای منفی و تقویت رفتارهای مدنی- تحصیلی می‌شود (گل پرور، ۱۳۹۴)؛ چنانچه دریافت رفتارهای عادلانه با خودداری از خشونت و درگیر شدن بیشتر در فعالیت‌های فوق‌برنامه

1. fairness review

(Resh & Sabbagh, 2017)، خستگی هیجانی، بی‌علاقگی تحصیلی و ناکارآمدی تحصیلی مرتبط است (مرزوقی و همکاران، ۱۳۹۲)؛ و حساسیت به عدالت با پذیرش و انجام رفتارهای فریبکارانه‌ی تحصیلی رابطه معنادار دارد چنانچه دریافت رفتارهای ناعادلانه با تقلب (Lemons & Seaton, 2011) و مدرسه گریزی (Ishak & Fin, 2013) مرتبط است و روش‌های نمره‌دهی ناعادلانه منجر به خطر افزایش ترک تحصیل دانشجویان می‌شود (Burger & Grob, 2016).

نظریه سؤال- پاسخ یک رویکرد مدل‌سازی قدرتمند است که برای ارزشیابی ویژگی‌های روان‌سنجی پرسشنامه‌های پیمایشی با پاسخ‌های مقوله‌ای (ترتیبی و غیرترتیبی) مانند مقیاس لیکرت به کار می‌رود (Embretson & Reise, 2000). مدل پاسخ مدرج یک نوع مدل سؤال- پاسخ است که امکان برازش مدل‌های چندبعدی را فراهم می‌کنند و تعمیم‌یافته مدل‌های دوپارامتری لجستیک هستند (Embretson & Reise, 2000, Zanon et al., 2016). در مدل پاسخ مدرج احتمال اینکه یک آزمودنی با توانایی θ یک مقوله مشخص یا مقوله بالاتر از آن را در یک سؤال انتخاب کند از طریق معادله زیر به دست می‌آید:

$$P_{ik}^{*}(\theta)=\frac{\exp\left[a_i(\theta_j- b_{ik})\right]}{1+\exp[a_i(\theta_j-b_{ik}]} \qquad (۱)$$

این معادله به‌عنوان تابع ویژه مرزی[1] آیتم i برای مقوله k شناخته می‌شود. پارامتر a_i شیب تابع یا ضریب تمیز سؤال است که برای همه مقوله‌های یک سؤال، ثابت است. پارامتر b_{ik} که پارامتر آستانه[2] نامیده می‌شود به نقطه‌ای روی مقیاس توانایی اشاره می‌کند که احتمال پاسخ‌گویی به یک مقوله مشخص یا بالاتر از آن ۵۰ درصد است (Hambleton et al., 1991؛ zanon et al., 2016). مدل پاسخ مدرج، هر سؤال با k مقوله را به‌عنوان k-1 سؤال دوارزشی در نظر می‌گیرد؛ به‌عنوان‌مثال اگر سؤال دارای ۵ مقوله ۰ تا ۴ باشد دوتایی‌های زیر مورد بررسی قرار می‌گیرد: ۰ در برابر ۱، ۲، ۳، ۴؛ ۰، ۱ در برابر ۲، ۳، ۴؛ ۰، ۱، ۲ در برابر ۳، ۴ و ۰، ۱، ۲، ۳ در برابر ۴. با توجه به معادله (۱) واضح است که $P_i^{*}(\theta)=۱$؛ ازاین‌رو پارامتر آستانه برای پایین‌ترین مقوله، برآورد نمی‌شود. در مورد بالاترین مقوله نیز با توجه به اینکه احتمال انتخاب مقولات بالاتر از آن برابر صفر است، مقدار به‌دست‌آمده برای $P_{i4}^{*}(\theta)$ دقیقاً

1. boundary characteristic function
2 . threshold parameter

برابر احتمال پاسخ به این مقوله خواهد بود. با توجه به معادله (۱) و توضیحات فوق احتمال این‌که یک فرد با توانایی θ مقوله k در سؤال i را انتخاب کند به معادله زیر به دست می‌آید:

$$P_{ik}(\theta)=P^{*}_{ik}(\theta)-P^{*}_{i(k+1)}(\theta) \qquad (۲)$$

احتمال پاسخگویی به ۵ مقوله به‌صورت زیر است:

$$P_{i0}(\theta) = 1 - P^{*}_{i1}(\theta)$$
$$P_{i1}(\theta) = P^{*}_{i1}(\theta) - P^{*}_{i2}(\theta)$$
$$P_{i2}(\theta) = P^{*}_{i2}(\theta) - P^{*}_{i3}(\theta)$$
$$P_{i3}(\theta) = P^{*}_{i3}(\theta) - P^{*}_{i4}(\theta)$$
$$P_{i4}(\theta) = P^{*}_{i4}(\theta) - 0$$

پژوهش‌هایی درزمینهٔ سنجش ادراک فراگیران و معلمان از عدالت در کلاس درس و روش‌های سنجش و نمره دهی (Chory et al., 2017; Čiuladienė & Račelytė, 2016; Alm, & Colnerud, 2015) پیامدهای ادراک عدالت در کلاس درس و سنجش و علل آن (ChoryAssad, 2002; Gordon & Fay, 2010) و ساخت ابزار برای ارزشیابی ادراک از سنجش کلاسی مانند پژوهش Brown (2006) که پرسشنامه ادراک معلمان از سنجش را مبتنی بر رساله دکترایش با عنوان «ادراک معلمان از سنجش» که با معلمان ابتدایی به‌کار رفته بود، تدوین کرد، انجام شده است؛ اما هیچ پژوهش خارجی که به شکل خاص ابزاری برای ارزشیابی سنجش کلاسی عادلانه تدوین کند، انجام نشده است. در ایران در حوزه ادراک دانشجویان از سنجش کلاسی، پژوهشی توسط Brown و همکاران (2014) با عنوان «ادراک دانشجویان دانشگاه‌های ایران درباره‌ی سنجش: کاربرد سنجش برای بهبود خود» انجام شده است که پژوهشی کمی مبتنی بر پرسشنامه ادراک فراگیران از سنجش است و روی چهار عامل مختلف بهبود، تأثیر، مرتبط نبودن و اسنادهای بیرونی تمرکز می‌کند. سایر پژوهش‌ها چالش‌های ارزشیابی دانشجویان را مورد بررسی قرار داده‌اند که می‌توانیم به پژوهش‌های سراجی و همکاران (۱۳۹۲)، قنبری و همکاران (۱۳۹۴) و قنبری و نیک‌خواه (۱۳۹۵) اشاره کنیم.

تنها پژوهشی که به‌طور خاص موضوع عدالت در سنجش کلاسی را مبتنی بر ادراک دانشجویان و اساتید ایرانی از عدالت مورد بررسی قرار داده است پژوهشی است که توسط بنی اسدی و همکاران (۱۴۰۰، در دست چاپ) انجام شده است و در آن برای ارزشیابی کیفیت عدالت در سنجش کلاسی در دانشگاه روبریک اولیه‌ای شامل ۲۰ ملاک سنجش که

برای هر ملاک توصیفگرهایی در چهار سطح عالی، خوب، متوسط و ضعیف تعریف شده است را تدوین کردند که پس از بررسی ویژگی‌های روان‌سنجی روبریک با کاربرد نظریه کلاسیک آزمون و حذف ملاک‌های ضعیف، ابزار نهایی مطلوبی شامل ۱۷ ملاک که بر اساس نتایج تحلیل عاملی اکتشافی و تأییدی شامل سه عامل عدالت رویه‌ای، ماهیت سنجش و عدالت رویه‌ای بود معرفی کردند.

با توجه به اینکه تنها ابزار ساخته‌شده برای ارزشیابی سنجش کلاسی عادلانه، روبریک سنجش کلاسی عادلانه (بنی اسدی و همکاران، ۱۴۰۰، در دست چاپ) تنها با استفاده از نظریه کلاسیک آزمون مورد بررسی قرارگرفته است و از منظر نظریه سؤال- پاسخ که سؤال مدار است ویژگی‌های روان‌سنجی آن بررسی نشده است و با عنایت به اینکه سؤال‌های روبریک سنجش کلاسی عادلانه چندارزشی با مقوله‌های پاسخ ترتیبی هستند این پژوهش به دنبال کاربرد مدل پاسخ مدرج در برآورد پارامترهای سؤال‌ها، مقوله‌های پاسخ و بررسی ویژگی‌های روبریک است؛ به این منظور به این سؤال‌ها پاسخ داده خواهد شد: ۱) کدام‌یک از مدل‌های پاسخ مدرج (یک‌بعدی یا چندبعدی) با داده‌ها برازش دارد؟ ۲) برآورد پارامترهای سؤال‌ها و مقوله‌های پاسخ و شاخص برازش سؤال‌ها بر اساس مدل مدرجِ انتخاب‌شده چگونه است؟

روش

نمونه این پژوهش ۵۱۱ دانشجوی دانشگاه تهران بودند که به روش نمونه‌گیری در دسترس انتخاب شدند و به روبریک سنجش کلاسی عادلانه پاسخ دادند. این تعداد با توجه به اینکه Reise and Yu (1990) حداقل ۵۰۰ آزمودنی را برای برآورد مناسب در مدل‌های پاسخ مدرج ضروری می‌دانند (Ayala, 2013) و همچنین Jiang و همکاران (2016) نیز در مطالعه‌ای که برای اندازه‌های نمونه مختلف، تعداد سؤال‌های متفاوت و همبستگی بین ابعاد به‌منظور برآورد پارامترها در مدل پاسخ مدرج چندبعدی انجام دادند، نشان دادند که برای اغلب موارد موردبررسی، اندازه نمونه برابر ۵۰۰ برآورد دقیق پارامترها را موجب خواهد شد و افزایش نمونه بیش از ۱۰۰۰ تا دقت برآورد پارامترهای مدل پاسخ مدرج چندبعدی را افزایش نمی‌دهد؛ مناسب است. ابزار گردآوری داده‌ها روبریک سنجش کلاسی عادلانه (بنی اسدی و همکاران، ۱۴۰۰، در دست چاپ) بود. این روبریک شامل ۱۷ ملاک است و هر

ملاک در چهار سطح عالی، خوب، متوسط و ضعیف با توصیفگرهایی مختص هر سطح تعریف شده است.

ابتدا باید بررسی شود که آیا مفروضه تک‌بعدی بودن برقرار است و امکان استفاده از مدل تک‌بعدی پاسخ مدرج وجود دارد. روش‌های مختلفی مانند مقیاس بندی چندبعدی[1]، تحلیل خوشه‌بندی سلسه‌مراتبی[2]، تحلیل موازی و ارزشیابی ابعاد برای تعیین صفات مشارکت‌کننده[3] به‌منظور بررسی تک‌بعدی بودن داده‌ها مورداستفاده قرار می‌گیرد (Zhang, 2013). در این پژوهش روش تحلیل موازی با استفاده از نرم‌افزار SPSS و روش دیتکت با استفاده از بسته sirt نرم‌افزار R برای بررسی این مفروضه استفاده شد. تحلیل موازی، فنی برای تعیین تعداد عامل‌های باقیمانده در تحلیل عاملی است که توسط Horn (1965) ارائه شده است. هورن این روش را به‌عنوان روشی جایگزین برای اصلاح مشکلات روش نمودار اسکری کتل مطرح کرده است. تحلیل موازی تلاش می‌کند تا بر محدودیت‌های معیار کایزر؛ یعنی بیش برآورد نظم ماتریس به علت خطای نمونه‌گیری غلبه کند (Glorfeld, 1995 به نقل از حجازی و همکاران، ۱۳۹۲) در یک ماتریس جامعه، ارزش ویژه برای متغیرهای تصادفی یا دوبه‌دو ناهمبسته، مساوی یک است اما در یک نمونه محدود خطای نمونه‌گیری و تورش حداقل مربعات منجر می‌شود که ارزش ویژه بزرگ‌تر از یک و یا کمتر به دست آید (Horn, 1965 به نقل از حجازی و همکاران، ۱۳۹۲). این بدان معنی است که برای نمونه‌های محدود برخی عوامل با ارزش ویژه بالاتر از یک ممکن است به‌عنوان نتیجه خطای نمونه‌گیری رخ دهد. تحلیل موازی اثر خطای نمونه‌گیری را تعدیل و بنابراین بر خلاف معیار کایزر که مبتنی بر جامعه است مبتنی بر نمونه است (Zwick & Velicer, 1986 به نقل از حجازی و همکاران، ۱۳۹۲). منطق تحلیل موازی این است که عامل‌های غیر بدیهی از داده‌های واقعی با یک ساختار عاملی معتبر باید ارزش ویژه بالاتر از عامل‌های مشتق شده از داده‌های تصادفی با تعداد مساوی حجم نمونه و تعداد متغیرهای مساوی داشته باشد (Lautenschlager, 1989). دیتکت سنجش ابعاد را بر اساس کوواریانس شرطی انجام می‌دهد و Sijtsma (2004) نشان داد که روش‌هایی که کوواریانس شرطی را به کار می‌برند در یافتن ساختارهای مشابه نسبت به کوواریانس غیرشرطی موفق‌تر عمل می‌کنند (Zhang, 2013). این شاخص با استفاده از بسته sirt

1. Multidimensional scaling (MDS)
2. Hierarchical cluster analysis (HCA)
3. dimensionality evaluation to enumerate contributing traits (DETECT)

نرم‌افزار R و تابع conf.detect به دست می‌آید. این تابع آماره دیتکت را برای پاسخ‌های دوارزشی و آماره polyDETECT را برای پاسخ‌های سؤال چندارزشی محاسبه می‌کند (Robitzsch & Robitzsch, 2020). خروجی این تابع سه شاخص ارزشیابی ابعاد برای تعیین صفات مشارکت‌کننده، شاخص ساختار ساده تقریبی[1] و شاخص نسبت[2] است (Zhang, 2007). مقادیر بحرانی برای این سه شاخص به ترتیب ۱، ۰/۲۵ و ۰/۳۶ است؛ به این معنی که اگر شاخص ارزشیابی ابعاد برای تعیین صفات مشارکت‌کننده کمتر از یک و نزدیک صفر باشد داده‌ها اساساً تک‌بعدی خواهند بود و مقادیر بزرگ‌تر از یک چندبعدی بودن قوی[3] را نشان می‌دهد همچنین مقادیر بزرگ‌تر از ۰/۲۵ برای شاخص ساختار ساده تقریبی و بیشتر از ۰/۳۶ برای شاخص نسبت انحراف اساسی از تک‌بعدی بودن داده‌ها را نشان می‌دهند (Chalmers et al., 2015).

در این پژوهش بسط چندبعدی مدل پاسخ مدرج (Samejima, 1969, 1997) برای برآورد و مقایسه مدل تک‌بعدی و چندبعدی به‌کار برده شد. این مرحله با استفاده از نرم‌افزار R و بسته نظریه سؤال- پاسخ چندبعدی[4] انجام شد. این بسته برای برآورد پارامترهای نظریه سؤال- پاسخ چندبعدی با استفاده از روش بیشینه درستنمایی[5] به وجود آمده است (Chalmers, 2012). برای برآورد پارامترهای مدل‌ها به توصیه Chalmers (2012) الگوریتم حداکثر انتظار[6] برای مدل تک‌بعدی و الگوریتم MetropolisHastings برای مدل چندبعدی استفاده شد. پارامترهای دشواری و تمیز برای مدلی که برازش بهتری با داده‌ها داشتند محاسبه شدند.

پارامتر تمیز نشان می‌دهد که تا چه اندازه سؤال، اساتیدی که در سطوح بالا و پایین سنجش کلاسی عادلانه قرار دارند را از یکدیگر تمیز می‌دهد. اگر این پارامتر بالا باشد در این صورت سؤال اطلاعات فراوانی درباره تفاوت‌های سنجش کلاسی عادلانه بین اساتید فراهم می‌کند و اگر پایین باشد اطلاعات زیادی در مورد تفاوت اساتید ازنظر سنجش کلاسی عادلانه فراهم نمی‌کند و ممکن است سؤال نیازمند اصلاح باشد و در مواردی نیز ممکن است به‌طورکلی حذف شود. بر اساس دیدگاه DeAyala (2013) سؤالاتی که شیب یا ضریب

1. approximate simple structure index (ASSI)
2. ratio index
3. Strong multidimensionality
4. multidimensional item response theory package (mirt)
5. maximum-likelihood
6. Expectation Maximization

تمیز بالای ۰/۸ دارند سؤالات مناسبی برای تمیز افراد دارای سطح بالا و افراد دارای سطح پایین صفت مورد سنجش هستند. پارامتر دشواری نشان می‌دهد که سطح سنجش کلاسی عادلانه اساتید چقدر باید باشد تا احتمال انتخاب یک مقوله پاسخ مشخص یا بالاتر از آن توسط دانشجویان ۰/۵ باشد. سؤال‌های روبریک سنجش کلاسی عادلانه شامل ۴ سطح عالی، خوب، متوسط و ضعیف است؛ بنابراین برای هر سؤال چهار پارامتر آستانه (دشواری) برآورد می‌شود.

برای سنجش برازش دو مدل و مقایسه آن‌ها آماره‌های ملاک اطلاعات آکایک[1]، آکایک تصحیح‌شده[2]، ملاک اطلاعات بیزی[3] و ملاک اطلاعات بیزی تعدیل‌شده بر اساس حجم نمونه[4] که آماره‌هایی برای مقایسه مدل‌ها هستند محاسبه شدند. مدل‌ها با مقادیر پایین‌تر در ملاک اطلاعات یکسان، مدل‌هایی با برازش بهتر در نظر گرفته می‌شوند (Hilbe, 2016). همچنین آماره‌های M2 کاهش‌یافته[5] که در مقایسه با سایر آماره‌ها وقتی اندازه نمونه نسبتاً کوچک است بهتر عمل می‌کند (Cai & Hansen, 2013، Depaoli et al., 2018) و زمانی که سطح معناداری این آماره بیشتر از ۰/۰۵ است نشان‌دهنده برازش مدل است؛ شاخص ریشه دوم میانگین مربعات باقی‌مانده[6] که با مقادیر کمتر از ۰/۰۸ نشان‌دهنده برازش مدل است (Singh, 2016) و آزمون نسبت درستنمایی[7] که در آن لگاریتم درستنمایی[8] مدل با همه پیش‌بینی‌کننده‌ها از مدل با پیش‌بینی‌کننده‌های کمتر، کم می‌شود و نتیجه در ۲- ضرب می‌شود و درجه آزادی این آزمون تفاوت تعداد پارامترهای دو مدل است (Hilbe, 2016)؛ محاسبه شدند. برای بررسی برازش هر سؤال آماره $S\text{-}X^2$ تعمیم‌یافته محاسبه شد (Kang & Chen, 2011;Orlando & Thissen, 2000) و از جهت کنترل مقایسه‌های چندگانه، نتایج معناداری آماره $S\text{-}X^2$ برای نرخ کشف اشتباه[9] تعدیل شدند (Benjamini & Hachberg, 1995). سطح معناداری کمتر از ۰/۰۵ نشان‌دهنده عدم برازش سؤال است. برای بررسی استقلال موضعی که جایگزین مفروضه ناهمبسته بودن خطاها در نظریه کلاسیک

1. Akaike information criterion (AIC)
2. AIC corrected (AICc)
3. Bayesian information criterion (BIC)
4. sample size adjusted BIC (sa BIC)
5. reduced M2
6. Root Mean Square Error of Approximation
7. Likelihood ratio tests (LRT)
8. log-likelihood
9. false discovery rate (FDR)

است (Kim & Feldt, 2010) از شاخص G^2 LD که به‌وسیله Chen and Thissen (1997) معرفی شده است استفاده شد. مقادیر معنادار نشان می‌دهد که ممکن است بین سؤالات کوواریانس باقیمانده وجود داشته باشد.

یافته‌ها

بررسی تک‌بعدی بودن بدین صورت انجام شد.

همان‌طور که اشاره شد برای بررسی تعداد ابعاد داده‌ها دو روش تحلیل موازی و دیتکت استفاده شد. نتایج تحلیل موازی در جدول ۱ ارائه شده است.

جدول ۱. میانگین و صدک ۹۵ ارزش‌های ویژه داده‌های تصادفی و داده‌های واقعی

عامل	ارزش‌های ویژه واقعی	میانگین ارزش‌های ویژه داده‌های تصادفی	صدک ۹۵ ارزش‌های ویژه داده‌های تصادفی	درصد واریانس تبیین شده توسط هر عامل
۱	۵/۲۸۰	۱/۴۰۵	۱/۴۸۳	۲۶/۴۰۱
۲	۴/۰۶۶	۱/۳۳۵	۱/۳۸۳	۲۰/۳۲۸
۳	۲/۰۷۰	۱/۲۷۹	۱/۳۰۹	۱۰/۳۵۲
۴	۰/۹۲۲	۱/۲۳۰	۱/۲۶۹	۴/۶۰۹

جدول ۱ نشان می‌دهد که سه عامل به‌دست‌آمده از تحلیل داده‌های واقعی ارزش ویژه بالاتر از عامل‌های مشتق شده از داده‌های تصادفی دارند؛ بنابراین تحلیل موازی نشان‌دهنده استخراج ۳ عامل است. همچنین در خروجی روش دیتکت شاخص ارزشیابی ابعاد برای تعیین صفات مشارکت‌کننده،۲/۸؛ شاخص ساختار ساده تقریبی، ۰/۲۶ و شاخص نسبت، ۰/۳۶۰۵۳ به دست آمدند که چندبعدی بودن قوی داده‌ها را نشان می‌دهد.

برازش مدل

شاخص‌های برازش مدل در جدول ۲ ارائه شده است. آزمون نسبت درستنمایی با توجه به مقدار لگاریتم درستنمایی دو مدل که در این جدول آمده است به‌صورت زیر محاسبه می‌شود:

LRT= -۲ (-۷۶۴۹/۸۲۱+۷۰۵۲/۳۰۳) =۱۱۵۹/۰۳۶

جدول ۲. برازش مدل‌های سؤال- پاسخ

شاخص برازش	مدل تک‌بعدی	مدل سه‌بعدی
log-likelihood	-۷۶۴۹/۸۲۱	-۷۰۵۲/۳۰۳
M2 (df)	۱۵۱/۲ (۷۲)*	۹۳/۱۵ (۶۹)
RMSEA	۰/۱۱۹	۰/۰۶۸
AIC	۱۵۴۲۹/۶۴	۱۴۲۴۰/۶۱
AICc	۱۵۴۴۸/۹۲	۱۴۲۶۱/۸۴
BIC	۱۵۷۰۵/۰۱	۱۴۵۲۸/۶۸
saBIC	۱۵۴۹۸/۶۹	۱۴۳۱۲/۸۴

*: سطح معناداری کمتر از ۰/۰۵

این آماره دارای توزیع خی دو با درجه آزادی ۳ است که در سطح ۰/۰۰۰ معنادار است و نشان می‌دهد که مدل چندبعدی به‌طور معناداری نسبت به مدل تک‌بعدی برازش بهتری با داده‌ها دارد. آماره M۲ کاهش‌یافته برای مدل تک‌بعدی معنادار بود [p=۰/۰۳، M۲(۷۲)=۱۵۱/۲] که نشان می‌دهد این مدل با داده‌ها برازش ندارد. بااین‌حال این آماره برای مدل چندبعدی معنادار نبود بود [p=۰/۳۵، M۲(۶۹)=۱۵۱/۲] که نشان‌دهنده بازتاب خوب داده‌ها توسط این مدل است. شاخص ریشه دوم میانگین مربعات باقی‌مانده برای مدل چندبعدی کمتر از مدل تک‌بعدی و کوچک‌تر از مقدار ۰/۰۸ است که برازش خوب این مدل را نشان می‌دهد. تمامی ملاک‌های اطلاعات (آکایک، بیزی، آکایک تصحیح‌شده و بیزی تعدیل‌شده) نیز برای مدل چندبعدی کمتر از مدل تک‌بعدی است که بازتاب بهتر داده‌ها توسط این مدل نسبت به مدل تک‌بعدی را نشان می‌دهد. پایایی و برآورد پارامترها بدین صورت انجام شد:

در چارچوب نظریه سؤال- پاسخ، پایایی هر یک از زیرمقیاس‌های عدالت رویه‌ای، ماهیت سنجش و عدالت تعاملی به ترتیب ۰/۸۵، ۰/۶۹ و ۰/۶۳ به دست آمد. برآورد پارامترهای نظریه سؤال- پاسخ شامل ضریب تمیز و ضریب دشواری در جدول ۳ ارائه شده است. در این جدول a_1 پارامتر تمیز سؤال‌های عامل اول، a_2 پارامتر تمیز سؤال‌های عامل دوم و a_3 پارامتر تمیز سؤال‌های عامل سوم است. با توجه به اینکه هر سؤال ۴ مقوله دارد ۳ ضریب دشواری برای این ۴ مقوله وجود دارد.

جدول ۳. برآورد سؤال- پاسخ پارامترهای سؤال‌ها برای مدل پاسخ مدرج

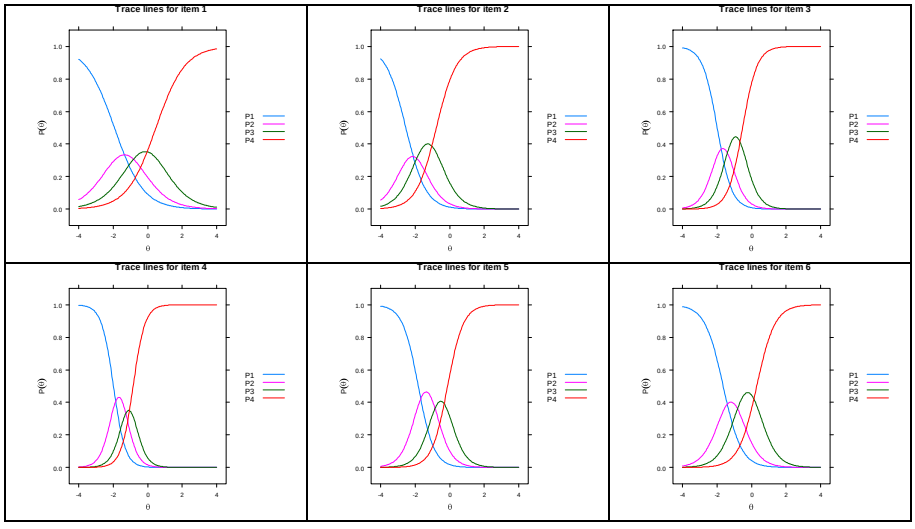
سؤال	تمیز (شیب)			دشواری (آستانه)		
	a_1	a_2	a_3	b_1	b_2	b_3
۷	۲/۳۸۲			-۱/۸۳۲	-۱/۰۱۱	-۰/۱۶۴
۳	۲/۳۵۹			-۱/۹۹۳	-۱/۳۲۸	-۰/۵۱۸
۵	۲/۱۹۴			-۱/۸۲۳	-۰/۹۰۷	-۰/۱۲۰
۱۶	۲/۱۳۳			-۲/۱۳۳	-۰/۷۶۷	۰/۱۸۲
۶	۱/۸۹۴			-۱/۶۴۶	-۰/۷۴۷	۰/۳۰۳
۲	۱/۷۱۷			-۲/۵۴۵	-۱/۷۶۲	-۰/۷۷۴
۸	۱/۵۶۶			-۲/۴۱۸	-۱/۱۷۲	۰/۰۹۷
۱	۱/۱۸۷			-۱/۹۳۲	-۰/۷۶۴	۰/۴۸۱
۹	۱/۱۳۶			-۳/۳۳۰	-۱/۳۶۴	۰/۴۴۰
۱۷	۱/۰۴۸			-۲/۷۷۳	-۰/۹۴۶	۰/۶۹۳
۱۱		۴/۰۸۴		-۱/۸۴۲	-۱/۳۶۶	-۰/۹۷۰
۱۰		۳/۶۳۳		-۱/۸۲۹	-۱/۳۵۲	-۰/۷۳۰
۱۲		۳/۲۵۸		-۲/۱۴۰	-۱/۴۴۱	-۰/۸۱۵
۴		۲/۸۹۶		-۱/۹۹۰	-۱/۳۵۲	-۰/۸۴۶
۱۳			۵/۸۰۲	-۲/۱۱۸	-۱/۴۸۱	-۱/۰۶۷
۱۵			۲/۹۹۳	-۱/۹۷۲	-۱/۶۲۵	-۱/۲۸۴
۱۴			۲/۸۷۳	-۲/۶۸۶	-۱/۸۴۵	-۱/۲۳۵

ضریب تمیز سؤال‌ها در دامنه ۱/۰۴۸ تا ۵/۸۰۲ قرار دارد که نشان می‌دهد همه سؤال‌ها در تمیز سطوح بالا و پایین سنجش کلاسی عادلانه به‌خوبی عمل می‌کنند. بسته mirt نرم‌افزار R تنها مقادیر عرض از مبدأ را محاسبه می‌کند که با فرمول $(-\frac{d}{a})$ مقادیر آستانه به دست می‌آیند. d مقدار عرض از مبدأ برای مقوله پاسخ مربوطه و a شیب سؤال است. به‌عنوان‌مثال سؤال ۷ در جدول ۵ شیبی برابر ۲/۳۸۲ دارد و نخستین عرض از مبدأ آن ۴/۳۶۶ است؛ بنابراین نخستین پارامتر دشواری برای این سؤال طبق فرمول ارائه‌شده برابر $\frac{-۴/۳۶۶}{۲/۳۸۲}$= -۱/۸۳۲ است. هر پارامتر دشواری نیز سطحی از صفت سنجش کلاسی عادلانه را نشان می‌دهد که در آن احتمال انتخاب مقوله مشخص یا مقولات بالاتر از آن در برابر مقولات پایین‌تر ۰/۵ است. به‌عنوان‌مثال برای سؤال ۶ پارامتر آستانه سوم برابر ۰/۳۰۳ است که نشان می‌دهد سطح صفت سنجش کلاسی عادلانه استادی باید ۰/۳۰۳ باشد تا دانشجویان با احتمال ۰/۵ مقوله عالی را در این

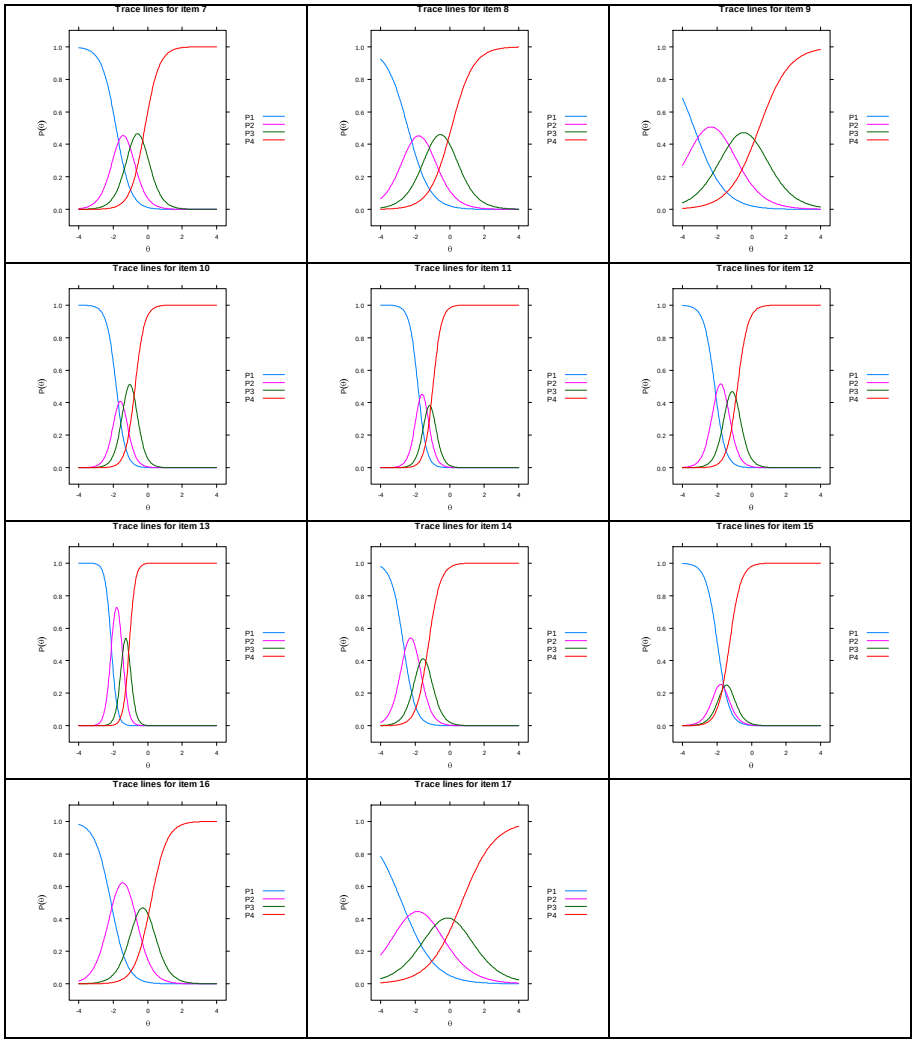
سؤال انتخاب کنند. همچنین می‌توان مقادیر آستانه را در بین سؤال‌ها مقایسه کرد. سؤال ۹ کمترین مقدار آستانه اول و سؤال ۶ بیشترین مقدار آستانه اول را دارد که این نشان می‌دهد تعداد افراد کمتری نخستین مقوله پاسخ سؤال ۹ را نسبت به نخستین مقوله پاسخ سؤال ۶ انتخاب کرده‌اند.

منحنی ویژه سؤال اطلاعاتی در سطح سؤال و درباره عملکرد مرتبط با هر مقوله پاسخ به دست می‌دهد. هر منحنی ویژه سؤال چهار خط اثر[1] مرتبط با مقوله‌های پاسخ سؤالات است و نشان‌دهنده احتمال انتخاب یک مقوله پاسخ سؤال با توجه به سطح معینی از صفت مکنون یا توانایی مکنون (در این پژوهش ابعاد سنجش کلاسی عادلانه) مشخص‌شده روی محور طول‌هاست. این نمودارها به‌طور مستقیم به ضرایب تمیز جدول ۳ پیوند می‌خورند به این معنی که سؤال‌ها با پارامتر تمیز بالاتر تمایل به ارائه اطلاعات در یک دامنه محدود دارند. به‌عنوان‌مثال سؤال ۱۳ دارای بیشترین مقدار ضریب تمیز است همان‌گونه که در شکل ۱ می‌بینید احتمال پاسخ به هر یک از مقوله‌های این سؤال در دامنه نسبتاً محدودی از مقادیر صفت مکنون روی محور طول‌ها متمرکز شده است. در مقابل سؤال ۱۷ دارای کمترین مقدار ضریب تمیز است و همان‌گونه که در شکل ۱ مشخص است احتمال مقوله‌های پاسخ این سؤال در طیف نسبتاً وسیعی از مقادیر صفت مکنون گسترده شده است.

شکل ۱. منحنی ویژه سؤال‌های روبریک سنجش کلاسی عادلانه



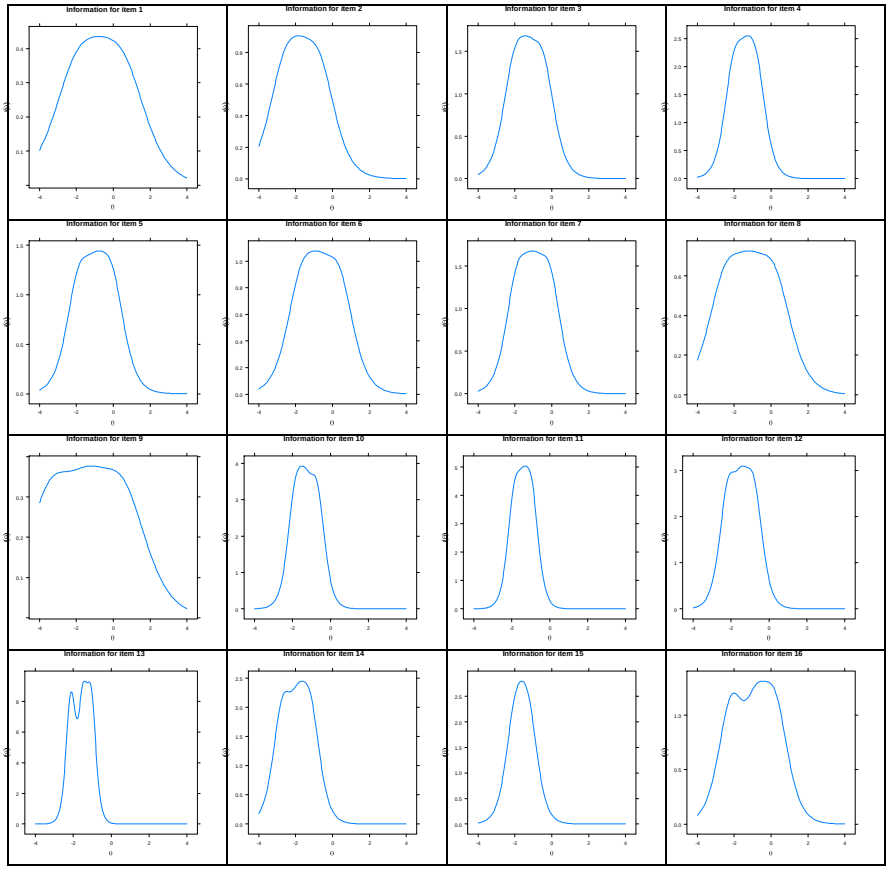
1. trace line

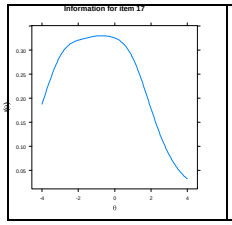


.تابع آگاهی میزان آگاهی را در بازه تتای ۴± برای هریک از گویه‌ها نشان می‌دهد مقدار آگاهی در یک سطح معین توانایی، برابر عکس واریانس است. زیاد بودن مقدار آگاهی به معنای آن است که می‌توان توانایی حقیقی امتحان شونده‌ای را که در آن سطح قرار دارد، با دقت برآورد کرد. به عبارتی همه برآوردها به گونه معقولی به نمره حقیقی وی نزدیک است؛ اما کم بودن مقدار آگاهی به معنای آن است که نمی‌توان توانایی را با دقت برآورد کرد و پراکندگی برآوردها پیرامون توانایی حقیقی بسیار زیاد است. در هریک از توابع محور افقی مقیاس نمره تتا را نشان می‌دهد و محور عمودی هم مقدار آگاهی را نشان

می‌دهد. توابع آگاهی سهم سؤال‌ها را در برآورد توانایی در طول پیوستار توانایی نشان می‌دهد (Hambleton et al., 1991) و وابسته به پارامتر شیب و پراکندگی دشواری مقولات پاسخ سؤال است. آگاهی که هر مقوله پاسخ سؤال ایجاد می‌کند با فرمول $I_i(\theta) = \frac{[P^*_{ik(\theta)}]^2}{P_{ik}(\theta)}$ محاسبه می‌شود و آگاهی سؤال مجموع آگاهی‌های تولیدشده توسط مقولات مختلف پاسخ برای یک سؤال است (DeAyala, 2013؛ Embretson & Reise, 2000).

شکل ۲. توابع آگاهی سؤال‌های روبریک





همان‌گونه که در شکل ۲ دیده می‌شود سؤال ۷ در عامل اول بالاترین آگاهی را در بازه توانایی ۲- تا ۱ ایجاد می‌کند، سؤال ۱۱ در عامل دوم بیشترین آگاهی را در بازه ۲- تا ۱ ایجاد می‌کند و سؤال ۱۳ در عامل سوم بیشترین آگاهی را در بازه ۲/۵- تا ۱ ایجاد می‌کند.

استقلال موضعی و برازش آیتم‌ها

بعد از کنترل نرخ کشف اشتباه آماره $S\text{-}X^2$ نشان داد که همه سؤال‌های روبریک از برازش خوبی برخوردارند نتایج کامل آماره برازش سؤال‌ها در جدول ۴ آمده است. سطوح معناداری برازش همه سؤال‌های روبریک را تائید می‌کند.

جدول ۴. آماره برازش سؤال‌ها

سؤال	$S\text{-}X^2$	df	p-value
۱	۵۲/۹۲۵	۴۹	۰/۷
۲	۳۷/۳۴۰	۳۳	۰/۷
۳	۳۵/۶۸۴	۳۴	۰/۸
۴	۴۴/۷۶۹	۳۷	۰/۷
۵	۴۶/۹۸۰	۴۱	۰/۷
۶	۴۸/۹۸۵	۴۴	۰/۷
۷	۳۸/۳۵۱	۴۰	۰/۸
۸	۴۷/۲۱۴	۴۲	۰/۷
۹	۴۰/۵۹۴	۳۴	۰/۷
۱۰	۳۸/۳۰۶	۳۵	۰/۷
۱۱	۳۶/۵۰۹	۳۳	۰/۷
۱۲	۳۲/۴۴۵	۲۹	۰/۷
۱۳	۲۳/۲۵۳	۲۳	۰/۸
۱۴	۲۵/۵۸۶	۲۱	۰/۷
۱۵	۲۳/۱۳۵	۲۲	۰/۸
۱۶	۳۶/۰۱۹	۳۵	۰/۸
۱۷	۵۰/۳۶۱	۴۲	۰/۵۲۶

بررسی شد. جدول کامل سطح معناداری LD G^2استقلال موضعی با استفاده از آماره برای ارزیابی استقلال موضعی بین آیتم‌ها در جدول ۵ ارائه شده است. جدول ۵ نشان می‌دهد که بین سؤالات کوواریانس باقی‌مانده وجود ندارد.

جدول ۵. مقادیر معناداری آماره G^2 استقلال موضعی

۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴	۱۵
–														
۰/۴۴۰	–													
۰/۶۵۱	۰/۵۸۴	–												
۰/۱۹۱	۰/۹۴۴	۰/۵۵۲	–											
۰/۵۳۰	۰/۵۵۹	۰/۰۵۵	۰/۲۴۹	–										
۰/۹۶۰	۰/۳۰۸	۰/۹۳۰	۰/۹۶۰	۰/۱۶۱	–									
۰/۶۵۰	۰/۳۲۲	۰/۴۳۳	۰/۶۸۳	۰/۹۷۰	۰/۰۵۵	–								
۰/۶۳۰	۰/۳۲۲	۰/۸۰۳	۰/۳۰۸	۰/۶۱۸	۰/۴۳۳	۰/۲۷۶	–							
۰/۳۹۰	۰/۴۳۳	۰/۴۳۳	۰/۷۳۱	۰/۷۳۱	۰/۳۲۵	۰/۶۳۳	۰/۹۵۰	–						
۰/۶۷۰	۰/۹۷۳	۰/۵۸۴	۰/۷۱۱	۰/۳۰۸	۰/۶۸۳	۰/۹۳۳	۰/۶۳۵	۰/۴۲۱	–					
۰/۷۶۰	۰/۸۰۲	۰/۹۲۹	۰/۹۵۰	۰/۵۳۴	۰/۴۳۳	۰/۴۲۱	۰/۷۴۴	۰/۵۸۹	۰/۹۸۰	–				
۰/۳۱۰	۰/۷۴۸	۰/۶۸۳	۰/۶۳۵	۰/۴۳۳	۰/۰۵۱	۰/۸۲۵	۰/۳۰۸	۰/۰۵۵	۰/۹۳۰	۰/۰۵۵	–			
	۰/۹۲۹	۰/۵۲۴	۰/۷۴۴	۰/۶۸۳	۰/۴۳۳	۰/۶۵۰	۰/۶۸۳	۰/۴۳۳	۰/۴۳۳	۰/۴۳۳	۰/۲۷۶	–		
	۰/۷۴۶	۰/۱۶۴	۰/۳۰۸	۰/۰۵۵	۰/۷۱۱	۰/۷۳۵	۰/۵۲۴	۰/۳۲۵	۰/۸۰۳	۰/۳۰۸	۰/۴۳۳	۰/۰۶۵	–	
۰/۸۳۰		۰/۵۲۶	۰/۶۲۴	۰/۹۳۰	۰/۳۱۹	۰/۸۰۳	۰/۱۶۴	۰/۶۸۳	۰/۲۴۹	۰/۵۳۴	۰/۰۵۵	۰/۹۶۰	۰/۰۵۱	–

۱	۰/۳۱۹	-	۰/۸۰۱	۰/۸۰۲	۰/۷۴۸	۰/۷۸۳	۰/۶۳۳	۰/۶۸۲	۰/۲۳۳	۰/۹۶۰	۰/۱۲۵	۰/۵۲۶	۰/۲۳۳	۰/۶۱۸	۰/۵۵۹	۰/۴۳۳	۰/۷۶۵	۰/۶۳۴	-	۱۶
	۰/۴۲۰		۰/۱۰۰	۰/۶۸۳	۰/۳۰۸	۰/۶۳۳	۰/۶۸۲	۰/۶۸۳	۰/۶۸۲	۰/۶۸۳	۰/۴۶۸	۰/۱۶۴	۰/۵۸۴	۰/۱۰۰	۰/۵۸۴	۰/۹۴۴	۰/۳۶۲			۱۷

بحث و نتیجه‌گیری

هدف این پژوهش بررسی ویژگی‌های روان‌سنجی روبریک سنجش کلاسی عادلانه در سطح سؤال و از منظر نظریه سؤال- پاسخ بود. در ابتدا مفروضه تک‌بعدی بودن برای برازش مدل متناسب با داده‌ها بررسی شد و چندبعدی بودن داده‌ها تائید شد؛ در چنین حالتی کاربرد مدل تک‌بعدی با داده‌های چندبعدی نتیجه مطلوبی به دست نمی‌دهد از طرفی استفاده از مدل‌های تک‌بعدی برای به دست آوردن پارامترهای نظریه سؤال- پاسخ در هر بعد جداگانه نیز منجر به مواردی همچون عدم برازش داده‌ها- مدل، از دست دادن نمرات مقیاس، کاهش اساسی در بهبود ضریب تمیز، کاهش دقت اندازه‌گیری، کاهش آگاهی سؤال‌ها و سوگیری نتایج خواهد شد (DeAyala, 1994)؛ بنابراین مدل چندبعدی پاسخ مدرج برای برازش با داده‌ها و به دست آوردن پارامترها انتخاب شد. یک مدل چندبعدی شامل خرده مقیاس‌های «عدالت رویه‌ای»، «عدالت تعاملی» و «ماهیت سنجش» نسبت به مدل تک‌بعدی به‌خوبی با داده‌ها برازش داشت؛ چنانچه تمام شاخص‌های برازش مدل شامل شاخص‌های لگاریتم درستنمایی، شاخص ریشه دوم میانگین مربعات باقی‌مانده، شاخص M2، نسبت درستنمایی و ملاک‌های اطلاعات مؤید این مطلب بودند و مسئله‌ای در مورد برازش سؤال‌ها و استقلال موضعی در هر بعد نیز وجود نداشت. پارامترهای تمیز سؤال‌ها نیز نشان داد که همه سؤال‌ها به‌خوبی بین اساتید با سطوح بالا و پایین سنجش کلاسی عادلانه تمیز قائل می‌شوند. بررسی تابع آگاهی سؤال‌ها در هر زیرمقیاس نشان داد که سؤال ۱۳ در زیرمقیاس عدالت تعاملی، سؤال ۱۱ در زیرمقیاس ماهیت سنجش و سؤال‌های ۳ و ۷ در زیرمقیاس عدالت رویه‌ای بیشترین آگاهی را در قسمتی از طیف توانایی ایجاد می‌کنند و دقت برآورد بالاتری دارند. درمجموع نتایج این پژوهش نشان‌دهنده آن است که روبریک ویژگی‌های روان‌سنجی مناسبی برای ارزشیابی عدالت در سنجش کلاسی اساتید دارد و کاربرد آن می‌تواند با شناسایی نقاط قوت و ضعف اساتید در حوزه عدالت سنجش به افزایش سواد سنجشی اساتید، افزایش کیفیت تدریس و

سنجش آن‌ها، افزایش بهزیستی دانشجویان و تقویت رفتارهای مدنی و تحصیلی آنان و تدوین دستورالعمل ملی سنجش کلاسی عادلانه کمک کند.

تعارض منافع

نویسندگان اذعان دارند که در این مقاله هیچ‌گونه تعارض منافعی وجود ندارد.

سپاسگزاری

پژوهش حاضر بدون همکاری مشارکت‌کنندگان امکان‌پذیر نبود؛ بدین‌وسیله از کلیه مشارکت‌کنندگان تقدیر و تشکر به عمل می‌آید.

منابع

ابطحی، سید حسین و ترابیان، محسن. (۱۳۸۹). بررسی تحقق اهداف آموزش عالی بر اساس سند چشم انداز بیست ساله کشور با روش فرایند تحلیل سلسله مراتبی (AHP). *فصلنامه پژوهش در نظام‌های آموزشی*، ۴(۸)، ۳۱-۶۰.

بنی اسدی، علی، صالحی، کیوان، خدایی، ابراهیم، باقری، خسرو و ایزانلو، بلال. (۱۴۰۰، در دست چاپ). ساخت روبریک برای ارزشیابی سنجش کلاسی عادلانه در دانشگاه. *اندازه‌گیری تربیتی*.

حجازی، الهه، نقش، زهرا و شیرزادی فرد، میثم. (۱۳۹۲). تحلیل موازی: روشی برای تعیین تعداد عامل‌ها. *اندازه‌گیری تربیتی*، ۱۵(۵)، ۱۰۵-۱۲۴.

سراجی، فرهاد، معروفی، یحیی و رازقی، طاهره. (۱۳۹۲). شناسایی چالش‌های ارزشیابی از آموخته‌های دانشجویان در نظام آموزش عالی ایران. *مطالعات اندازه گیری و ارزشیابی آموزشی*، ۴(۵)، ۳۳-۵۴.

قنبری، سیروس، اردلان، محمدرضا و کریمی، ایمان. (۱۳۹۴). تأثیر چالش‌های ارزشیابی آموخته‌های دانشجویان بر رویکرد مطالعه عمل فکورانه. *دوماهنامه علمی- پژوهشی راهبردهای آموزش در علوم پزشکی*، ۸(۲)، ۱۰۵-۱۱۳.

قنبری، مهدی و نیکخواه، محمد. (۱۳۹۵). فهم چالش‌های عنصر ارزشیابی برنامه درسی دانشجویان دوره کارشناسی رشته علوم تربیتی دانشگاه فرهنگیان: جستاری پدیدارشناسانه. *فصلنامه علمی- تخصصی آموزش پژوهی*، ۲(۷)، ۴۷-۶۲.

کاویانی، حسن و نصر، احمدرضا. (۱۳۹۵). سنتزپژوهی چالش‌های برنامه‌های درسی آموزش عالی کشور در دهه اخیر و راهکارهای پیش رو. *دوفصلنامه مطالعات برنامه درسی آموزش عالی*، ۷(۱۳)، ۷-۳۴.

گل پرور، محسن. (۱۳۹۴). نقش اخلاق تحصیلی در رابطه عدالت و بی عدالتی آموزشی با فریب کاری تحصیلی: مدل معادله ساختاری. *فصلنامه روانشناسی تربیتی*، ۱۱(۳۷)، ۵۱-۶۵.

مرزوقی، رحمت اله، حیدری، معصومه و حیدری، الهام. (۱۳۹۲). بررسی رابطه عدالت آموزشی با فرسودگی تحصیلی دانشجویان دانشگاه علوم بهزیستی و توانبخشی. *گام‌های توسعه در آموزش پزشکی*، ۱۰(۳)،۳۲۸-۳۳۴.

هلب، جی‌ام. (۲۰۱۶). *راهنمای عملی رگرسیون لجستیک*. ترجمه علی بنی اسدی و ابراهیم خدایی (۱۳۹۹)، تهران: جهاد دانشگاهی.

همبلتون، آرک، سوامیناتان، اچ و راجرز، اچ جی. (۱۹۹۱). *مبانی نظریه‌ی پرسش پاسخ*. ترجمه محمدرضا فلسفی نژاد (۱۳۸۹)، تهران: دانشگاه علامه طباطبایی.

References

Alm, F., & Colnerud, G. (2015). Teachers' experiences of unfair grading. *Educational Assessment*, *20*(2), 132–150.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300.

Brown, G. T. (2006). Secondary school students' conceptions of assessment: a survey of four schools. *Conceptions of Assessment and Feedback Project Report*, *5*.

Brown,G. T. L., Pishghadam, R., & Sadafian, S. (2014). Iranian university students' conceptions of assessment: Using assessment to self-improve. *Assessment Matters*, 5-33.

Burger, R., & Grob, M. (2016). Fairness and university dropout. The role of grading procedures in the development of dropout intentions.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of statistical Software*, *48*(1), 1-29.

Chalmers, P., Pritikin, J., Robitzsch, A., & Zoltak, M. (2015). *Package 'mirt'*. https://cran.r-project.org/web/packages/mirt/mirt.pdf

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265-289.

Chory-Assad, R. M. (2002). Classroom justice: Perceptions of fairness as a predictor of student motivation, learning, and aggression. *Communication Quarterly*, *50*(1), 58–77.

Chory, R. M., Horan, S. M., & Houser, M. L. (2017). Justice in the higher education classroom: Students' perceptions of unfairness and responses to instructors. *Innovative Higher Education*, *42*(4), 321–336.

Čiuladienė, G., & Račelytė, D. (2016). Perceived unfairness in teacher-student conflict situations: students' point of view. *Polish Journal of Applied Psychology*, *14*(1), 49–66.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.
DeLuca, C. (2012). Preparing teachers for the age of accountability: Toward a framework for assessment education. *Action in Teacher Education, 34*, 576-591.
Depaoli, S., Tiemensma, J., & Felt, J. M. (2018). Assessment of health surveys: Fitting a multidimensional graded response model. *Psychology, health & medicine*, *23*(sup1), 1299-1317.
Educational Testing Service. (2009). *ETS Guidelines for Fairness Review of Assessments.* Princeton: NJ: ETS.
Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Maheah.
Gordon, M. E., & Fay, C. H. (2010). The effects of grading and teaching practices on students' perceptions of grading fairness. *College Teaching*, *58*(3), 93–98.
Ishak, Z., & Fin, L. S. (2013). Truants' and Teachers' Behaviors in the Classroom. *Procedia - Social and Behavioral Sciences*, *103*.
Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in psychology*, *7*, 109.
Kang, T., & Chen, T. T. (2011). Performance of the generalized SX 2 item fit index for the graded response model. *Asia Pacific Education Review*, *12*(1), 89-96.
Kim, S., & Feldt, L. S. (2010). The estimation of the IRT reliability coefficient and its lower and upper bounds, with comparisons to CTT reliability statistics. *Asia Pacific Education Review*, *11*(2), 179-188.
Lautenschlager, G. J. (1989). A comparison of alternatives to conducting Monte Carlo analyses for determining parallel analysis criteria. *Multivariate behavioral research*, *24*(3), 365-395.
Lemons, M., & Seaton, J. (2011). Justice in the classroom: Does fairness determine student cheating behaviors? *Journal of Academic Administration in Higher Education*, *7*(1).
Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied psychological measurement*, *24*(1), 50-64.
Resh, N., & Sabbagh, C. (2017). Sense of justice in school and civic behavior. *Soc Psychol Educ*, *20*, 387-409.
Robitzsch, A., & Robitzsch, M. A. (2020). *Package 'sirt'*. https://cran.r-project.org/web/packages/sirt/sirt.pdf
Singh, K., Junnarkar, M., & Kaur, J. (2016). *Measures of positive psychology*. Development and Validation. Berlin: Springer.
Zanon, C., Hutz, C. S., Yoo, H. H., & Hambleton, R. K. (2016). An application of item response theory to psychological test development. *Psicologia: Reflexão e Crítica*, *29*.
Zhang, J. (2007). Conditional covariance theory and detect for polytomous items. *Psychometrika*, *72*(1), 69-91.
Zhang, J. (2013). A procedure for dimensionality analyses of response data from various test designs. *Psychometrika*, *78*(1), 37-58.

استناد به این مقاله: بنی اسدی، علی، صالحی، کیوان، خدایی، ابراهیم، باقری، خسرو و ایزانلو، بلال. (۱۴۰۱). روبریک سنجش کلاسی عادلانه: برازش مدل چندبعدی پاسخ مدرج. *فصلنامه اندازه‌گیری تربیتی*، ۱۳(۴۹)، ۳۱-۵۱.
doi: 10.22054/jem.2023.63328.2219