Educational Measurement
https://jem.atu.ac.ir/

ATU PRESS

Comparing the Results of the Likert Scales Item's Analysis in the Classical Test Theory and the Item Response Theory

Reihane Rahimi *

Corresponding Author, Assistant Professor, Department of Educational Science, Farhangian University, Tehran, Iran. E-mail: reyhanerahimi1367@gmail.com

Aso Mojtahedi

Ph.D. Student in Measurement & Assessment, Allameh Tabataba'i University, Tehran, Iran. E-mail: asomojtahedi@gmail.com

14(55), Spring 2024, 146-183
eISSN: 2470-6240 ISSN: 2252-004x
Article Type: Reserch Article
Received January 25, 2023 Received in revised form April 15, 2023
Published Online April 03, 2024 Accepted October 24, 2023

Abstract

The goal of this research is to explore the Likert scale questions using two distinct methods: Classical Test Theory and Item Response Theory. By comparing the results of these approaches, the study aims to address the question: "Do the outcomes from these two methodologies align, or do they contradict each other?" The research design followed a descriptive methodology and utilized secondary analysis techniques. The study population consisted of 977 junior high school students. After the data screening process, the final sample size for analyzing extraversion items was 783 students, 763 students for openness items, and 784 students for conscientiousness items. The research instruments were the three subscales of extraversion, openness, and conscientiousness from the Neo Personality Test. The statistical analysis yielded results indicating that a strong internal consistency among items enhanced the accuracy and validity of outcomes derived from the graded response model. However, when items exhibit low internal consistency, caution should be exercised, as the model may yield erroneous thresholds or discrimination coefficients (i.e., false negative or positive). Overall, combining multiple methods of statistical analysis can significantly contribute to more effective analysis and obtaining highly accurate results.

Keywords: Graded Response Model; Item Response Theory; Classical Test Theory; Internal Consistency; Threshold; Discrimination Coefficient

Cite this Article: Rahimi, R., & Mojtahedi, A. (2024). Comparing the Results of the Likert Scales Item's Analysis in the Classical Test Theory and the Item Response Theory. *Educational Measurement*, *14*(55), 146-183. https://doi.org/10.22054/jem.2023.71953.3431

Publisher: Allameh Tabataba'i University Press
DOI: https://doi.org/10.22054/jem.2023.71953.3431

1. Introduction

In the field of psychometrics, the study of the relationship between a continuous latent variable and a categorical observed variable is referred to as Item-Response Theory (IRT) (Ostini & Nering, 2006). Unlike the Classical Test Theory (CTT), which focuses on the overall score, IRT is significant as it emphasizes the analysis of items individually. Concentrating on the item level allows for the development, revision, and optimization of scales tailored to specific applications (Baker, 2001; De Ayala, 2013; Embretson & Reise, 2000; Hambleton et al., 1985). In this study, our focus is on exploring polytomous items with Likert scale from both the perspective of Item-Response Theory (IRT) and Classical Test Theory (CTT). The primary objective of this research is to compare the outcomes generated by the two approaches – Item-Response Theory (IRT) and Classical Test Theory (CTT) – and determine whether their results align or conflict with one another. A secondary yet practical aim is to introduce polytomous IRT models to researchers, aiming to facilitate a better comprehension and accurate interpretation of the concepts and outcomes associated with these models. Additionally, the study seeks to address the challenge of understanding polytomous models and their limited applications in a predominantly dichotomous item context.

2. Literature Review

The Graded Response Model (GRM), introduced by Muraki in 1990, is extensively employed in the analysis of Likert scale items. GRM is also known as the Reduced Graded Response Model (R-GRM) or the Parsimonious, Constrained, or Modified Model. The primary distinction between the Graded Response Model (GRM) and the reduced model, the Reduced Graded Response Model (R-GRM), lies in the estimation of discrimination indices. In the R-GRM, a single discrimination index is computed for all items with the assumption of equal discrimination indices across all items. In contrast, the GRM estimates a unique discrimination index for each individual item, allowing for more precise analysis. Both GRM and R-GRM are suitable for analyzing items on a Likert scale (Toland, 2014).

The Graded Response Model (GRM) extends the two-parameter model used for items with multiple ordinal response options. Unlike the two-parameter model, GRM assigns a unique slope parameter to each item, providing more nuanced insights. Furthermore, GRM

distinguishes itself from the two-parameter model in terms of the computation of discrimination index. In contrast to the two-parameter model, which estimates the difficulty parameter, Graded Response Model (GRM) instead calculates the threshold. For each item, the number of thresholds is equal to the total number of categories minus 1. In the Reduced Graded Response Model (R-GRM), a uniform discrimination parameter is estimated for every item. The primary advantage of the R-GRM (Reduced Graded Response Model) model is that it is more parsimonious, as it lessens the total number of parameters required for estimation. However, this simplification implies the assumption that the relationship between the latent trait and each item remains uniform.

3. Methodology

The participants in this research comprised 977 first-year high school students who had previously completed the Neo test. The intended analysis represents a secondary analysis since the data was already collected for another purpose. All available information was considered for analysis, and no additional data sampling was conducted. During the data screening process, certain participants with item nonresponse were excluded from analysis due to the limitations of item-response models. Separate screening was performed for each personality trait to ensure an accurate representation and reliable evaluation. The analysis of extraversion items was based on the data of 783 participants, while the analysis of openness to experience items was conducted using the information of 763 individuals. Similarly, the analysis of conscientiousness items was carried out utilizing the data of 784 subjects. The research was conducted using 3 subscale items of the NEO personality questionnaire, specifically focused on extraversion, openness to experience, and conscientiousness. These subscales were selected because they showcased varying Cronbach's alpha coefficients, permitting the researcher to effectively pursue the objectives of the study.

The alpha coefficient of the openness characteristic amounted to 0.234, indicative of very weak internal consistency amongst the items in this set. The alpha coefficient for extraversion reached 0.527, denoting an averageinternal consistency within the items in this category. In stark contrast, the alpha coefficient for conscientiousness was considerably higher, at 0.759, demonstrating high internal

consistency amongst the items within this category. The presented data analysis comprises 3 distinct sections, each delving into the following aspects: (1) the analysis of openness items, where the lowest internal consistency was found, (2) the analysis of extraversion items, with an average internal consistency, and finally (3) the analysis of conscientiousness items, which displayed the highest internal consistency. Each section incorporates descriptive information, along with CTT and IRT-based analyses.

Within the section dedicated to descriptive information, the following aspects are detailed: (1) the frequency percentage corresponding to each option response, (2) the average (mean) and dispersion (standard deviation) of each subscale, as well as (3) the mean for each item. In the context of dichotomous items, the mean can be considered an approximate representation of the difficulty index. Analysis within this section was conducted using SPSS software. The examination of items via CTT involves calculating the internal consistency coefficient, with Cronbach's alpha serving as the metric for this purpose. Additionally, the correlation between each item and the overall score (which can be taken as the discrimination index for dichotomous items) is reported, alongside the alpha coefficient in the event of removing each individual item. In the section pertaining to the analysis of IRT on Likert scale items, the Graded Response Model (GRM) proposed by Simjima was utilized. Before carrying out the analysis, a comparison between the fit of the GRM and the R-GRM was performed using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and LRT index, the full report of which can be found in Table 1. For all three subscales, the GRM was deemed superior to its reduced model, exhibiting a more favorable fit with the data. Therefore, all subsequent analyzes

4. Conclusion

In summary, it is evident that the presence of weak items significantly enhances the likelihood of errors and can lead to erroneous conclusions when employing IRT-based analyzes. Our observations align with this, as seen in the case with openness, where items with very low alpha coefficients and weak or even negative correlations between items were found.

Furthermore, the threshold levels encountered in the items of this collection reach seemingly implausible values such as 11 and 10, which

deviates from the typical range reported by Baker (2001). Generally, latent trait scores and threshold parameters are usually confined between -2 and 2, and even values spanning from -3 to 3 may not be entirely unreasonable. Values beyond this range, however, are atypical and might indicate potential issues with the items or less meaningful responses.

Therefore, employing a scale with suboptimal psychometric qualities (characterized by poor accuracy across or at certain locations along the continuum of the latent trait for items) can compromise the interpretation of potential score inferences, and if utilized as an output variable in a statistical analysis, it may lead to a decrease in the validity of statistical conclusion inferences (Kang & Waller, 2005). The research findings of Zamanpour et al. (2018) further support the notion that if the unidimensionality assumption is rigidly adhered to, the items determined to be suitable in both analyses will be identical. Consequently, if all the assumptions are fulfilled and the items exhibit satisfactory psychometric properties, there should be minimal difference between the two types of analyses. Nevertheless, IRT analyzes offer more thorough and nuanced information. For instance, in the analysis of conscientiousness items, the superiority of IRT models was confirmed. This superiority was attributed to the fact that information regarding the information functions of these items was unattainable through CTT analysis alone. In the analysis of conscientiousness items, while the CTT approach demonstrated all appropriate conditions and there was no evidence of any issues, IRT analysis revealed that these items were incapable of detecting individuals with conscientiousness levels exceeding 2 and failed to provide informative data.

IRT analysis should be preceded by a check to ensure that there is sufficient frequency in each category and that all response options have been adequately selected by the participants of the sample. This study observed that in problematic items, most individuals tended to select specific options, which led to unrealistic ICC values. While there are no strict and explicit guidelines regarding the criterion of adequacy, having more responses within each category can enhance the accuracy of the estimation of item parameters and facilitate the evaluation of the usefulness of the response categories.

When insufficient numbers are employed in the response categories, it may be necessary to aggregate them together to create a reduced

response category system. This approach enhances the accuracy and stability of the item parameter estimation (Toland, 2014). In line with the findings of this research, it would be beneficial to utilize both theoretical approaches for examining the items of each test. This approach guarantees that the weaknesses of each theory are explored and effectively addressed, leading to a comprehensive test evaluation. Furthermore, it is recommended that future studies compare various IRT models with CTT analyses, in order to objectively observe and assess the strengths and limitations of each approach from a practical standpoint.

اندازه‌گیری تربیتی
https://jem.atu.ac.ir/

دوره ۱۴، شماره ۵۵، بهار ۱۴۰۳، ۱۴۶-۱۸۳
ISSN: 2252-004x
eISSN: 2470-6240

نوع مقاله: مقاله پژوهشی
تاریخ دریافت: ۱۴۰۱/۱۱/۰۵؛ تاریخ بازنگری: ۱۴۰۲/۰۱/۲۶
تاریخ پذیرش: ۱۴۰۲/۰۸/۰۲؛ تاریخ انتشار: ۱۴۰۳/۰۱/۱۵

مقایسه نتایج تحلیل سؤالات طیف لیکرت با نظریه کلاسیک آزمون و نظریه سؤال پاسخ

ریحانه رحیمی *
نویسنده مسئول، استادیار گروه آموزش علوم تربیتی، دانشگاه فرهنگیان، تهران، ایران. رایانامه: reyhanerahimi1367@gmail.com

ئاسو مجتهدی
دانشجوی دکتری رشته سنجش و اندازه‌گیری، دانشگاه علامه طباطبائی، تهران، ایران. رایانامه: asomojtahedi@gmail.com

چکیده

هدف این مطالعه بررسی سؤالات طیف لیکرت با دو رویکرد نظریه کلاسیک آزمون و نظریه سؤال پاسخ و قرار دادن نتایج این دو رویکرد در کنار یکدیگر و پاسخ به این سؤال است که «آیا نتایج این دو رویکرد همسو بوده یا اینکه در تضاد با یکدیگرند؟». روش پژوهش از نوع توصیفی و از حیث تحلیل داده‌ها یک روش تحلیل ثانویه است. ابزار پژوهش ۳ خرده مقیاس برونگرایی، گشودگی و مسئولیت‌پذیری آزمون شخصیت نئو فرم کوتاه (NEO-FFI) بود. آزمودنی‌های موردمطالعه در این پژوهش ۹۷۷ نفر از دانش‌آموزان مقطع متوسطه اول بودند که با حذف افراد در مرحله غربالگری داده‌ها، درنهایت تحلیل سؤالات برونگرایی بر روی اطلاعات ۷۸۳ نفر آزمودنی، تحلیل سؤالات گشودگی بر روی اطلاعات ۷۶۳ نفر آزمودنی و تحلیل سؤالات مسئولیت‌پذیری نیز بر روی اطلاعات ۷۸۴ نفر آزمودنی انجام شد. نتایج تحلیل‌های آماری نشان داد که هرچه همسانی درونی سؤالات بالاتر باشد نتایج دقیق‌تر و معتبرتری از مدل مدرج پاسخ به دست می‌آید و برای سؤالات با همسانی درونی پایین از طریق این مدل باید بسیار محتاط عمل کرد، زیرا ممکن است آستانه سؤالات را به‌صورت وارونه نشان داده یا ضریب تشخیص سؤالات را منفی یا مثبت کاذب نشان دهد. درمجموع کاربرد هم‌زمان روش‌های مختلف تجزیه‌وتحلیل آماری به تحلیل بهتر و رسیدن به نتایج دقیق‌تر کمک خواهد کرد.

کلیدواژه‌ها: مدل مدرج پاسخ، نظریه سؤال پاسخ، نظریه کلاسیک آزمون، همسانی درونی، آستانه، ضریب تشخیص

استناد به این مقاله: رحیمی، ریحانه و مجتهدی، ئاسو. (۱۴۰۳). مقایسه نتایج تحلیل سؤالات طیف لیکرت با نظریه کلاسیک آزمون و نظریه سؤال پاسخ. *فصلنامه اندازه‌گیری تربیتی*، ۱۴(۵۵)، ۱۴۶-۱۸۳.
https://doi.org/10.22054/jem.2023.71953.3431

ناشر: دانشگاه علامه طباطبائی

مقدمه

ساختن هر مقیاس یا پرسشنامه‌ای که به لحاظ روان‌سنجی برای اندازه‌گیری یک صفت مکنون مناسب و کارآمد باشد، مستلزم همکاری تیمی از کارشناسان محتوا، متخصصان موضوعی و روان‌سنجی یا کارشناسان سنجش است (Reeve & Fayers, 2005). وظیفه کارشناس حوزه سنجش در این زمینه استفاده از روش‌های مختلف و معتبر در جهت تأیید یا رد مقیاس موردنظر یا تعدیل و اصلاح سؤالات آن برای رسیدن به مقیاسی معتبر برای اندازه‌گیری هر چه دقیق‌تر آن صفت مکنون از طریق نمودهای آشکار آن است؛ اما اگر طبق نظر Messick (1989) روایی[1] را فرایندی بی‌پایان بدانیم که همواره باید بررسی و شواهدی در جهت تأیید آن جمع‌آوری شود، آیا وظیفه یک متخصص بعد از ساخت آزمون تمام می‌شود؟ بر اساس این دیدگاه همواره باید در جهت یافتن شواهدی برای روایی تلاش کرد. تیلور[2] حتی به این مقدار نیز رضایت نداده و معتقد است که روایی و رواسازی[3] در همه وجوه و مراحل تحقیق وجود داشته و مستلزم پیگیری و بررسی بوده و منحصر به ساخت آزمون نمی‌شود. ازنظر تیلور هر عاملی که باعث ایجاد شبهه در ادعاهای مربوط به سنجش و پژوهش شود، تهدیدی علیه روایی است. او راه چاره برای مقابله با این تهدیدها را رواسازی می‌داند و رواسازی را زیر سؤال بردن ادعاها در ۴ حوزه روایی درونی[4]، روایی بیرونی[5]، روایی نتیجه‌گیری آماری[6] و روایی سازه[7] معرفی می‌کند (Taylor, 2013، ترجمه یونسی، ۱۳۹۸). اگرچه روایی نتیجه‌گیری آماری را تیلور بیشتر از منظر روش‌های آزمایشی موردبحث قرار می‌دهد، اما این روایی می‌تواند در سایر پژوهش‌ها نیز مدنظر قرار گیرد. با پیشرفت علم آمار و رشد روزافزون مدل‌های آماری، هر پژوهشگری در بررسی مسئله و فرضیهٔ پژوهشی‌اش با چندین روش مختلف برای رسیدن به مقصود روبه‌رو است که بعضاً باعث سردرگمی و تردید او می‌شود. یکی از این سردرگمی و تردیدها خاص حوزه روان‌سنجی است. هر پژوهشگر روان‌سنجی در مسیر پژوهش خود بیشتر با دو رویکرد نظریه کلاسیک آزمون[8] و نظریه سؤال

---

1. validity
2. Taylor, C. S.
3. validation
4. internal validity
5. external validity
6. statistical conclusion validity
7. construct validity
8. Classical Test Theory (CTT)

پاسخ[1] روبه‌رو است و می‌تواند با انتخاب یکی از این رویکردها یا حتی تلفیق دو رویکرد به بررسی مسئله خود بپردازد.

نظریه کلاسیک آزمون اولین نظریه منسجم در حیطه اندازه‌گیری بود که در اوایل قرن بیستم مطرح شد (Courville, 2004). در حقیقت تلاش‌های Spearman (1904) منجر به گسترش این نظریه شد و حدود نیم‌قرن تنها نظریه در حوزه آزمون‌سازی بود. نظریه کلاسیک آزمون، مدلی از خطای اندازه‌گیری بر اساس ضریب همبستگی است. ضریب همبستگی که توسط اسپیرمن بیان شد برای توضیح خطای اندازه‌گیری از دو مفهوم همبستگی واقعی و همبستگی مشاهده‌شده کمک می‌گیرد. ضریب همبستگی و نظریه کلاسیک آزمون هر دو بر این فرض استوار هستند که میانگین به‌دست‌آمده از همه اندازه‌گیری‌های ممکن با اندازه واقعی در جامعه برابر هستند (Cochran, 1997, cited in Courville, 2004). فرض بنیادی نظریه کلاسیک این است که نمره مشاهده‌شده فرد، ترکیبی از نمره واقعی و نمره خطا است (Lord & Novick, 1968). در دهه ۱۹۷۰ و اوایل ۱۹۸۰ نظریه کلاسیک تنها ابزار روان‌سنجی بود که در سنجش پایایی آزمون‌ها و نیز در طراحی آزمون‌های استاندارد به کار می‌رفت (Hambleton & Jones, 1993)؛ اما با رشد روزافزون کامپیوترها به‌تدریج امکان استفاده از نظریه سؤال پاسخ بیشتر شد و بعدازآن به یک چهارچوب اندازه‌گیری مهم تبدیل گردید (Hambleton & Swaminathan, 2013).

در ادبیات روان‌سنجی تحلیل رابطه بین متغیر پنهان پیوسته با متغیر مشاهده‌شده طبقه‌ای را نظریه سؤال پاسخ می‌نامند (Ostini & Nering, 2006). از منظر تئوری، نظریه سؤال پاسخ توانست بر ضعف‌های اساسی نظریه کلاسیک غلبه کرده و مدل‌های آن توانایی این را دارند که ویژگی‌های سؤال‌ها را مستقل از نمونه آزمودنی‌ها و نیز توانایی آزمودنی‌ها را مستقل از نمونه سؤال‌های اجراشده، محاسبه کنند. این خصیصه نامتغیر بودن[2] ویژگی‌های آماری سؤال و آزمودنی، به‌طور نظری در IRT نشان داده می‌شود (Hambleton & Swaminathan, 2013). درواقع نظریه سؤال پاسخ مجموعه‌ای از مدل‌های آماری را شامل می‌شود که از طریق مدل‌سازی ریاضی بین متغیر پنهان پیوسته فرد و ویژگی‌های سؤال در پی برآورد احتمال پاسخ درست به سؤال هستند. اهمیت نظریه سؤال پاسخ این است که برخلاف نظریه کلاسیک که بر نمره کل تأکید دارد بر تجزیه‌وتحلیل در سطح سؤال متمرکز است. تمرکز

1. Item Response Theory (IRT)
2. invariance

در سطح سؤال امکان ابداع، بازنگری و بهینه‌سازی مقیاس‌ها را برای کاربردهای خاص فراهم می‌کند (Baker, 2001; De Ayala, 2013; Embretson & Reise, 2000; Hambleton et al., 1985). البته نظریه کلاسیک نیز روش‌های خاص خودش مثل بررسی ضریب تشخیص و دشواری سؤال یا تأثیر سؤال در ضریب همسانی درونی را برای بررسی سؤالات دارد.

سؤال اساسی که در اینجا مطرح می‌شود این است که اگر نظریه سؤال پاسخ در جهت رفع مشکلات نظریه کلاسیک مطرح شد و توسعه یافت چرا هنوز با حجم زیادی از پژوهش‌ها روبه‌رو هستیم که از روش‌های نظریه کلاسیک در جهت بررسی آزمون‌ها استفاده می‌کنند؟ آیا علت، ناکارآمدی این نظریه جدید است یا اینکه دشواری بیش‌ازحد کار با مدل‌های این نظریه و دشواری تفسیر و درک آن توسط کاربران و پژوهشگران باعث ایجاد چنین وضعیتی شده است؟ یا اینکه نظریه کلاسیک با تمام مشکلاتش توان پاسخ‌گویی به نیازهای آزمون‌سازی را دارد؟ حتی در برخی از پژوهش‌ها مشاهده می‌شود که در صورت استفاده از نظریه سؤال پاسخ از نظریه کلاسیک نیز کمک گرفته شده است. برای مثال گاهی برای تعیین تعداد ابعاد یا تشخیص تک‌بعدی بودن آزمون از تحلیل عاملی استفاده شده است (Toland, 2014; Jeong & Lee, 2016).

با توجه به اهمیت روایی نتایج آماری که در ابتدای بحث مطرح شد و با در نظر گرفتن سؤالاتی که درباره چرایی مغفول ماندن نظریه سؤال پاسخ در میدان عمل، با وجود همه تعاریفی که از آن در نوشتارهای نظری شده است، در این مطالعه بر آن شدیم که به بررسی سؤالات چند ارزشی با طیف لیکرت از دو منظر IRT و CTT بپردازیم. اولین هدف این مطالعه، قرار دادن نتایج این دو رویکرد در کنار یکدیگر و پاسخ به این سؤال که آیا نتایج این دو رویکرد همسو بوده یا اینکه ممکن است نتایج در تضاد با یکدیگر باشند. دومین هدف از این مطالعه که بیشتر جنبه کاربردی دارد، آشناکردن پژوهشگران با مدل‌های IRT چند ارزشی و تلاش در جهت درک و تفسیر درست مفاهیم و نتایج این مدل‌ها است، زیرا تا حدی درک مدل‌های چند ارزشی دشوار بوده و کاربردهای IRT به همان سؤال‌های دوارزشی محدود شده است.

با توجه به اهداف ذکرشده در این پژوهش سؤالات آزمون نئو فرم کوتاه[1] که دارای طیف لیکرت هستند استفاده شد. شاید ساده‌ترین و واضح‌ترین دلیل برای توسعه مدل‌های

1. NEO-FFI

IRT چند ارزشی این واقعیت باشد که معمولاً در اندازه‌گیری روان‌شناختی کاربردی از سؤالات چند ارزشی استفاده می‌شود؛ بنابراین نظریه سؤال پاسخ برای اینکه یک رویکرد اندازه‌گیری جامع باشد، راهی جز ارائه روش‌های مناسب برای مدل‌سازی این داده‌ها نداشت تا بتواند در بررسی آزمون‌های شخصیت و نگرش سنج نیز کاربرد داشته باشد (Ostini & Nering, 2006). از طرف دیگر معمولاً برای سنجش روایی از تحلیل عامل تأییدی استفاده می‌شود، درصورتی‌که تحلیل عاملی برای سؤال‌های با پاسخ پیوسته مناسب است درصورتی‌که طیف لیکرت دارای پاسخ‌های طبقه‌ای بوده و مدل پاسخ مدرج در نظریه سؤال پاسخ برای این منظور مناسب‌تر است (Jeong & Lee, 2016). به همین خاطر قبل از ورود به روش کار و یافته‌ها، ابتدا توضیحاتی در جهت آشنایی خوانندگان با مدل‌های پاسخ مدرج و پاسخ مدرج کاهش‌یافته در بخش بعدی ارائه شده است.

پیشینه پژوهش

مدل‌های اولیه نظریه سؤال پاسخ در رابطه با سؤال‌های دوارزشی (دارای جواب درست یا غلط) به وجود آمدند؛ ولی به‌مرورزمان مدل‌های دیگری برای بررسی سؤال‌هایی با بیش از دو گزینه که گزینه درست و غلط درباره آن‌ها صدق نمی‌کرد معرفی شدند. یکی از این مدل‌ها، مدل پاسخ مدرج[1] است که توسط Samejima (1969) معرفی شده و معمولاً بیشترین کاربرد را در تحلیل سؤالاتی با طیف لیکرت دارد. مدل دیگری که در ادامه این مدل توسط Muraki (1990) معرفی شده مدل پاسخ مدرج کاهش‌یافته[2] (همچنین معروف به محدودشده[3]، تعدیل‌شده[4] یا مقرون‌به‌صرفه[5]) است. تفاوت مدل GRM با R-GRM در این است که در مدل کاهش‌یافته برای تمامی سؤالات یک ضریب تشخیص برآورد می‌گردد و فرض بر برابری ضرایب تشخیص است، درصورتی‌که در مدل اصلی برای هر سؤال ضریب تشخیص جداگانه‌ای برآورد می‌گردد. هر دو این مدل‌ها برای بررسی سؤالاتی با طیف لیکرت مناسب هستند (Toland, 2014).

1. Graded Response Model (GRM)
2. Reduced Graded Response Model (R-GRM)
3. constrained
4. modified
5. parsimonious

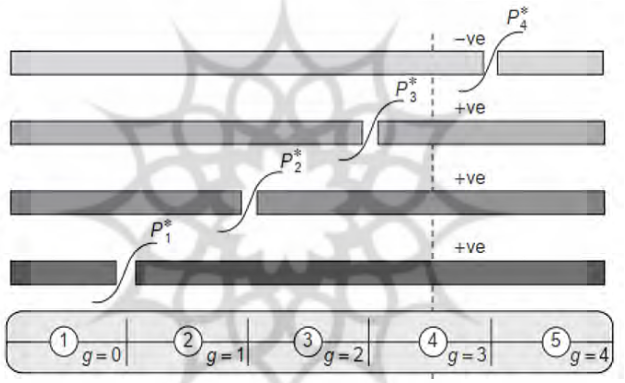
مدل GRM بسطی از مدل دو پارامتری است که برای سؤالات با بیش از دو گزینه که گزینه‌های آن نیز از نوع ترتیبی است، کاربرد دارد. مدل GRM یک پارامتر شیب منحصربه‌فرد برای هر سؤال ارائه می‌دهد و به لحاظ ضریب تشخیص تفاوتی با مدل دو پارامتری ندارد، اما به‌جای پارامتر مکانی یا همان دشواری در مدل دو پارامتری، این مدل به برآورد آستانه[1] می‌پردازد. تعداد آستانه‌های هر سؤال برابر است با تعداد گزینه‌ها آن سؤال منهای یک، مثلاً در آزمون نئو که دارای ۵ گزینه است، ۴ آستانه برای هر سؤال برآورد می‌گردد؛ بنابراین در یک آزمون ۱۲ سؤالی با ۵ گزینه، ۱۲ پارامتر تشخیص و ۴۸ آستانه برآورد می‌گردد که روی‌هم‌رفته ۶۰ پارامتر برای چنین آزمونی برآورد می‌شود؛ اما اگر برای چنین آزمونی از مدل R-GRM استفاده شود کل پارامترهای برآورد شده برابر با ۴۹ پارامتر است؛ زیرا تنها یک پارامتر تشخیص برای تمامی سؤالات برآورد می‌شود. مزیت مدل R-GRM این است که صرفه‌جوتر است، زیرا تعداد پارامترهای موردنیاز برای تخمین را کاهش می‌دهد، اما با این ساده‌سازی، رابطه بین صفت نهفته و هر سؤال یکسان فرض می‌شود.

یکی از مهم‌ترین موارد در کاربرد چنین مدل‌هایی فهم و تفسیر خروجی‌های مدل است. آستانه یکی از پارامترهای مهم و اصلی مدل GRM و R-GRM است که بعضاً دانشجویان و پژوهشگران در تفسیر و فهم آن دچار مشکل می‌شوند. آشنایی با مفهوم پارامتر مکانی یا همان شیب در مدل‌های دوارزشی تا حد زیادی می‌تواند در درک مفهوم آستانه کمک‌کننده باشد. پارامتر مکانی برابر است با آن میزان از توانایی (صفت مکنون) که احتمال انتخاب پاسخ درست برابر با احتمال پاسخ غلط باشد، یعنی فردی که دارای توانایی مساوی با پارامتر مکانی است، شانس درست جواب دادنش به آن سؤال برابر ۰/۵ است. حال در این مدل‌ها چون با بیش از دو گزینه روبرو هستیم، هر بار گزینه‌ها به دودسته تقسیم می‌شوند؛ به این شکل که در ابتدا گزینه اول (گزینه با امتیاز کمتر) در مقابل گزینه‌های سطح بالاتر قرار گرفته و برآورد می‌شود که چه میزانی از آن صفت مکنون لازم است تا احتمال انتخاب گزینه اول برابر با احتمال انتخاب مجموعه گزینه‌های بالاتر باشد، این میزان از صفت برآورد شده همان آستانه اول است. در برآورد آستانه دوم، دو گزینه اول (با امتیاز کمتر) در یک دسته و گزینه‌های بالاتر در یک دسته قرار گرفته و باز میزانی از صفت مکنون که لازم است تا احتمال انتخاب این دو طبقه برابر با یکدیگر شوند، برآورد شده و معادل با آستانه دوم تعریف می‌شود. این دسته‌بندی‌های دوتایی تا جایی ادامه پیدا می‌کند که تمامی گزینه‌های ابتدایی

1. threshold

در یک دسته و گزینه انتهایی در دسته دیگر قرار گرفته و آستانه آخر برآورد می‌شود؛ بنابراین آستانه به لحاظ مفهوم تفاوتی با پارامتر مکانی نداشته و حتی می‌توان این‌گونه تعبیر کرد که پارامتر مکانی همان آستانه در سؤالات دوارزشی بوده و چون دو گزینه بیشتر وجود ندارد و تعداد آستانه نیز برابر با تعداد گزینه‌ها منهای یک است، فقط یک آستانه برآورد شده و نام این یک آستانه را پارامتر مکانی یا دشواری گذاشته‌اند. در شکل زیر که برگرفته از کتاب «مدل‌های نظریه سؤال پاسخ چند ارزشی» تألیف Ostini and Nering (2006) است مفهوم آستانه به‌خوبی تصویرسازی شده است. در تصویر زیر سؤالی با ۵ گزینه (g) و ۴ آستانه (P) و طبقه‌بندی‌های ذکرشده و مرز بین طبقات قابل‌مشاهده است.

شکل ۱. نمایش گرافیکی مدل مدرج پاسخ



در پاراگراف قبلی راجع به آستانه و تفسیر آن بحث شد. به‌غیراز آستانه، تفسیر و درک نمودارها یا منحنی‌های حاصل از مدل‌های چند ارزشی نیز دارای اهمیت است. اولین مشکل در شناسایی این منحنی‌ها، نبود یک اسم مشخص و واحد برای نامیدن آن‌هاست. در هر کتاب و مقاله‌ای بنا بر سلیقه نویسنده نامی برای این منحنی‌ها در نظر گرفته شده است. یکی از منحنی‌های استخراج‌شده از این دو مدل منحنی ویژه عملیاتی[1] است. منحنی ویژه عملیاتی به‌نوعی معادل همان ICC در سؤالات دوارزشی است با این تفاوت که ICC برای هر سؤال یک منحنی لاجیت تجمعی در نظر می‌گیرد ولی OCC برای هر آستانه یک منحنی لاجیت تجمعی ارائه می‌دهد. در این منحنی محور افقی متعلق به صفت مکنون و محور عمودی مربوط به احتمال است و برای هر آستانه یک منحنی لاجیت تجمعی رسم شده است. این

1. Operational Characteristic Curve (OCC)

منحنی با نام‌های دیگر ازجمله منحنی ویژه کرانی[1] یا تابع ویژه عملیاتی[2] نیز معرفی می‌شود؛ علاوه بر این ممکن است نام‌های دیگری نیز برای این منحنی وجود داشته باشد، مهم آن است که بدانید چنین توابعی دارای منحنی لاجیت تجمعی به تعداد آستانه‌های سؤال هستند. منحنی دیگر که بسیار در این مدل‌ها کاربرد دارند منحنی ویژه طبقات[3] هستند که احتمال انتخاب هر گزینه را نشان می‌دهند. نکته مهم در رابطه با این منحنی این است که در هر سطحی از صفت مکنون، مجموع احتمالات گزینه‌ها برابر با ۱ می‌شود. اگر هر گزینه در یک سؤال به‌درستی عمل کند، باید در هر سطحی از صفت مکنون یک گزینه بیشترین احتمال را برای انتخاب‌شدن داشته باید و ترتیب قرارگرفتن گزینه‌ها بر روی پیوستار صفت مکنون از کوچک به بزرگ باشد؛ یعنی هرچه در پیوستار صفت به سمت بالا می‌رویم احتمال انتخاب گزینه‌های با امتیاز بیشتر افزایش یابد. این منحنی با نام‌های دیگر مثل تابع پاسخ عملیاتی[4] یا تابع پاسخ طبقه[5] نیز معرفی می‌شود (Ostini & Nering, 2006).

روش

آزمودنی‌های موردمطالعه در این پژوهش ۹۷۷ نفر از دانش‌آموزان مقطع متوسطه اول بودند که آزمون شخصیت نئو برای آن‌ها با هدف ارزیابی ویژگی‌های شخصیتی‌شان اجرا شده بود؛ بنابراین چون این داده‌ها قبلاً و با هدف دیگری جمع‌آوری شده‌اند، تحلیل موردنظر از حیث داده‌ها یک تحلیل ثانویه محسوب می‌شود. تمامی این اطلاعات برای تحلیل در نظر گرفته شده و نمونه‌گیری از داده‌ها به عمل نیامد؛ البته در جریان غربالگری داده‌ها، به دلیل محدودیت مدل‌های سؤال پاسخ، آزمودنی‌هایی که سؤال بدون پاسخ داشتند از تحلیل کنار گذاشته شدند. غربالگری برای هر ویژگی شخصیتی به‌صورت جداگانه انجام شد، برای مثال در بررسی ۱۲ سؤال مربوط به ویژگی برونگرایی افرادی از مجموعه آزمودنی‌ها حذف شدند که حداقل به یکی از ۱۲ سؤال مربوط به برونگرایی پاسخ نداده بودند و این رویه برای دو ویژگی دیگر نیز به همین منوال صورت گرفت. درنهایت تحلیل سؤالات برونگرایی بر روی

1. Boundary Characteristic Curve (BCC)
2. Operational Characteristic Function (OCF)
3. Category Characteristic Curve (CCC)
4. Operating Response Function (ORF)
5. Category Operating Response Function (CORF)

اطلاعات ۷۸۳ نفر آزمودنی، تحلیل سؤالات گشودگی بر روی اطلاعات ۷۶۳ نفر آزمودنی و تحلیل سؤالات مسئولیت‌پذیری نیز بر روی اطلاعات ۷۸۴ نفر آزمودنی انجام شد.

ابزار مورداستفاده در این پژوهش همان‌طور که قبلاً ذکر شد، سؤالات مربوط به ۳ خرده مقیاس پرسشنامه شخصیت نئو، شامل برون‌گرایی، گشودگی و مسئولیت‌پذیری بود. این ۳ خرده مقیاس ازآن‌جهت انتخاب شدند که ضریب آلفای کرونباخ متفاوتی داشته و می‌توانستند پژوهشگر را در رسیدن به هدف این پژوهش یاری کنند. ضریب آلفا ویژگی گشودگی برابر ۰/۲۳۴ و نشان‌دهنده همسانی درونی بسیار ضعیف بین سؤالات این مجموعه است. ضریب آلفا ویژگی برون‌گرایی برابر با ۰/۵۲۷ و حاکی از همسانی درونی متوسط بین سؤالات این مجموعه است. ضریب آلفا ویژگی مسئولیت‌پذیری برابر با ۰/۷۵۹ است که نشان از همسانی درونی مطلوب بین سؤالات این مجموعه است. قبل از انجام هرگونه تجزیه‌وتحلیلی، سؤالات معکوس هر خرده مقیاس مطابق با دستورالعمل نمره‌گذاری معکوس شده و سپس تحلیل‌های نهایی بر روی آن‌ها انجام شد.

تجزیه‌وتحلیل داده‌ها در ۳ بخش سؤالات گشودگی (با کمترین میزان همسانی درونی)، سؤالات برون‌گرایی (با میزان همسانی درونی متوسط) و سؤالات مسئولیت‌پذیری (با بیشترین میزان همسانی درونی)، ارائه شده است که هرکدام شامل اطلاعات توصیفی، تحلیل مبتنی بر نظریه کلاسیک آزمون و تحلیل مبتنی بر نظریه سؤال پاسخ می‌شوند. در بخش اطلاعات توصیفی درصد فراوانی انتخاب هر گزینه در هر سؤال، میانگین و انحراف استاندارد هر خرده مقیاس و میانگین مربوط به هر سؤال که همان ضریب مقبولیت سؤال است و معادل دشواری در سؤالات دوارزشی است، گزارش شده است. تحلیل‌های این بخش از طریق نرم‌افزار SPSS انجام شد. تحلیل مبتنی بر نظریه کلاسیک آزمون به بررسی ضریب همسانی درونی سؤالات پرداخته و از ضریب آلفا کرونباخ برای رسیدن به این هدف استفاده کرده است. همچنین همبستگی بین هر سؤال با نمره کل آزمون که معادل ضریب تشخیص است و ضریب آلفا در صورت حذف هر سؤال گزارش شده است. در بخش مربوط به تحلیل‌های نظریه سؤال پاسخ با توجه به طیف لیکرت سؤالات، از مدل پاسخ مدرج پیشنهادی سیمجیما استفاده شد. قبل از انجام تحلیل، برازش مدل پاسخ مدرج و مدل پاسخ مدرج کاهش‌یافته با استفاده از شاخص ضریب آکائیک[1] و شاخص ضریب بیزی[2] و شاخص LRT مورد مقایسه گرفتند

1. Akaike Information Criterion (AIC)
2. Bayesian Information Criterion (BIC)

که گزارش کامل آن در جدول ۱ آمده است و برای هر ۳ خرده مقیاس مدل پاسخ مدرج بر مدل کاهش‌یافته آن برتری داشته و برازش بهتری با داده‌ها نشان داد، بنابراین همه تحلیل‌ها بر اساس مدل پاسخ مدرج که شامل آستانه‌ها و ضرایب تشخیص متفاوت برای هر سؤال است، انجام شد. همچنین تابع آگاهی کلی و تابع آگاهی مربوط به هر سؤال به‌علاوه منحنی‌های ویژه عملیاتی و منحنی آگاهی مربوط به هر سؤال در این بخش گزارش شده است. تمامی تحلیل‌های این بخش از طریق نرم‌افزار RStudio و با استفاده از پکیج "ltm" انجام شد. کدهای دستوری اجرا شده در پیوست آمده است.

جدول ۱. شاخص‌های برازش دو مدل پاسخ مدرج و پاسخ مدرج کاهش‌یافته

ویژگی	مدل	AIC	BIC	LRT	p.value
گشودگی	GRM-R	۲۶۵۲۰/۴۳	۲۶۷۴۷/۶۶	۲۶۳/۱۴	<۰,۰۰۱
	GRM	۲۶۲۷۹/۲۹	۲۶۵۵۷/۵۲		
برون‌گرایی	GRM-R	۲۶۴۱۲/۰۸	۲۶۶۴۰/۵۷	۳۰۰/۷۸	<۰,۰۰۱
	GRM	۲۶۱۳۳/۳	۲۶۴۱۳/۰۹		
مسئولیت‌پذیری	GRM-R	۲۶۴۱۲/۰۸	۲۶۶۴۰/۵۷	۵۸/۶۷	<۰,۰۰۱
	GRM	۲۳۱۱۴/۴	۲۳۳۹۴/۲۶		

بر اساس اطلاعات مندرج در جدول بالا، شاخص ضریب آکائیک و شاخص ضریب بیزی برای هر سه ویژگی در مدل پاسخ مدرج کوچک‌تر از مدل پاسخ مدرج کاهش‌یافته است و درنتیجه مدل پاسخ مدرج نسبت به مدل کاهش برازش بهتری دارد. چون به‌طورمعمول مدل‌های پیچیده‌تر برازش بهتری نسبت به مدل‌های کاهش‌یافته و ساده‌تر نشان می‌دهند (Sorre et al., 2016) به این دو شاخص اکتفا نشده و از آزمون LRT که برای مقایسه بین دو مدل که یکی در دیگری آشیانه[1] کرده، کاربرد دارد، استفاده شد. نتیجه آزمون LRT نشان داد که بین دو مدل به لحاظ آماری تفاوت وجود داشته و برتری مدل پاسخ مدرج برحسب شانس نبوده است؛ بنابراین تمامی تحلیل‌های گزارش شده در بخش یافته‌ها مبتنی بر مدل پاسخ مدرج است.

1. nested

یافته‌ها

در این بخش اطلاعات به‌دست‌آمده از تجزیه‌وتحلیل‌های آماری گزارش شده است. تحلیل‌های انجام‌شده برای مجموعه سؤالات هر خرده مقیاس به‌طور جداگانه ارائه شده است؛ به‌این‌ترتیب که ابتدا تحلیل‌های مربوط به ویژگی گشودگی، سپس اطلاعات مربوط به ویژگی برون‌گرایی و درنهایت تحلیل مربوط به ویژگی مسئولیت‌پذیری آمده است. درواقع ابتدا سؤالات با کمترین همسانی درونی (سؤالات گشودگی) سپس سؤالات با همسانی درونی متوسط (سؤالات برون‌گرایی) و درنهایت سؤالات با همسانی درونی مطلوب (سؤالات مسئولیت‌پذیری) گزارش شده‌اند. لازم به ذکر است که به خاطر رعایت اختصار، در جداول و نمودارها به‌جای ارائه صورت سؤال شماره سؤالات درج گردیده است. برای آگاهی از صورت سؤالات به پیوست مراجعه شود.

گشودگی: این ویژگی سومین ویژگی شخصیتی موردمطالعه در آزمون شخصیت نئو است که با ۱۲ سؤال اندازه‌گیری می‌شود. اندازه نمونه موردمطالعه در این ویژگی برابر با ۷۶۳ نفر است که میانگین نمرات گشودگی آن‌ها برابر با ۲۶/۱۹ و انحراف استاندارد آن برابر با ۴/۷۲ است. کمترین نمره گشودگی برابر با ۹ و بیشترین آن برابر با ۴۰ است. در جدول زیر درصد فراوانی انتخاب هر گزینه توسط افراد، میانگین مربوط به هر سؤال (ضریب مقبولیت)، همبستگی سؤال با نمره کل (ضریب تشخیص بر اساس نظریه کلاسیک آزمون) و آلفا در صورت حذف سؤال گزارش شده است.

جدول ۲. فراوانی درصدی انتخاب هر گزینه، میانگین و همبستگی هر سؤال با نمره کل و ضریب آلفا در صورت حذف سؤال

سؤال	گزینه ۱	گزینه ۲	گزینه ۳	گزینه ٤	گزینه ٥	میانگین	همبستگی	آلفا
۱	۳۹/۱	۲۳/۳	۱۲/۸	۱۳/٥	۱۱/۳	۱/۳٥	-۰/۰۱٦	۰/۲٦۷
۲	٤٥/۹	۲۷/٤	۱٤/٥	۷/۱	٥/۱	۰/۹۸	-۰/۰٥۰	۰/۲۷۲
۳	۳/٤	٤/۲	۹/٤	۲۸/٦	٥٤/٤	۳/۲٦	۰/۱۱٥	۰/۲۰۳
٤	۱۱/۱	۱۰/٦	۲٤	۱۸/۱	۳٦/۲	۲/٥۸	۰/۱۳۷	۰/۱۸٥
٥	۸/۳	۱۱/۱	۲۰/٦	۳۰/۸	۲۹/۲	۲/٦۲	۰/۱٥۲	۰/۱۸۱
٦	۱٥/٦	۱۸/۹	۲٤/۹	۲۳/۷	۱٦/۹	۲/۰۷	۰/۰۲۹	۰/۲٤۱
۷	۱۰/۲	۱٦/۳	۳۸/۱	۲۲	۱۳/٤	۲/۱۲	۰/۰٥۳	۰/۲۲۸
۸	۳۲/۹	۳٥/٥	۱۹/۸	۷/۳	٤/٥	۱/۱٥	-۰/۱۲۳	۰/۲۹۸

سؤال	گزینه ۱	گزینه ۲	گزینه ۳	گزینه ٤	گزینه ٥	میانگین	همبستگی	آلفا
۹	۸/۷	۱٥/۲	۲۸/۳	۲۹/٦	۱۸/۲	۲/۳٤	۰/۲۰۰	۰/۱٥۹
۱۰	٦/۲	۸/٤	۲۱/٦	۲۹/۸	۳٤/۱	۲/۷۷	۰/۱۸۳	۰/۱٦۸
۱۱	۳/۳	۷/٦	۱۹/٤	۳۱/٥	۳۸/۳	۲/۹٤	۰/۱۲۲	۰/۱۹۹
۱۲	۱۳/٦	۲۰/۳	۳۰/۷	۲۱/۲	۱٤/۲	۲/۰۲	۰/۱۱۳	۰/۲۰۰

بر اساس درصد فراوانی‌های گزارش شده در جدول بالا، در سؤالات ۱، ۲ و ۸ بیش از ۶۰ درصد افراد دو گزینه ابتدایی را انتخاب کرده‌اند و در سؤالات ۳، ۱۰ و ۱۱ بیش از ۶۰ درصد افراد دو گزینه پایانی را برگزیده‌اند. بیشترین ضریب مقبولیت برای سؤال ۳ و برابر با ۳/۲۶ و کمترین آن مربوط به سؤال ۲ و برابر با ۰/۹۸ است. با توجه به محتوای این دو سؤال و در نظر گرفتن این نکته که سؤالات معکوس قبل از انجام هر تحلیلی در جهت سازه گشودگی نمره‌گذاری شده‌اند، می‌توان گفت که بیشتر افراد موردمطالعه در صورت یافتن راه درست انجام یک کار، مدام آن را تکرار می‌کنند و پدیده‌های طبیعی و هنری نیز آن‌ها را مبهوت می‌کند.

سؤالات ۱، ۲ و ۸ دارای ضریب تشخیص منفی بوده و با سؤالات دیگر مربوط به گشودگی رابطه منفی نشان داده‌اند. از طرف دیگر با توجه به ضریب آلفای کرونباخ مربوط به ۱۲ سؤال گشودگی که برابر با ۰/۲۳۴ است، در صورت حذف سؤالات ۱، ۲، ۶ و ۸ ضریب آلفا افزایش می‌یابد. این نتیجه نشان می‌دهد که این ۴ سؤال تهدیدی برای همسانی درونی سؤالات این مجموعه هستند. در ادامه تحلیل مبتنی بر نظریه سؤال پاسخ که با مدل "GRM" انجام‌شده گزارش شده است. در جدول زیر ضریب تشخیص و آستانه‌های مربوط به هر سؤال گزارش شده است.

جدول ۳. ضریب تشخیص و آستانه‌های مربوط به هر سؤال

سؤال	ضریب تشخیص	آستانه ۱	آستانه ۲	آستانه ۳	آستانه ٤
۱	۰/۲۷۹	-۱/٥۹۲	۱/۸۸٦	٤/۰۸٤	۰/۲۷۹
۲	۰/٤۳٥	-۰/۳۹٤	۲/٤۱٥	٤/٦۹٤	٦/۸۹۲
۳	-۱/۰٥٥	۳/٦۲۹	۲/۷٦۷	۱/۷۹۳	۰/۲۱۰
٤	-۰/٥٥۳	۳/۹۹۰	۲/٥۱۸	۰/٤۰۰	-۱/۰٥۳
٥	-۰/٦۸۳	۳/۷۸٤	۲/۲۸٥	۰/٦۷۷	-۱/٤۱٥
٦	-۰/۱۸۸	۹/۰۷۲	۳/٤۹٤	-۱/۹۸۹	-۸/٥۱۲

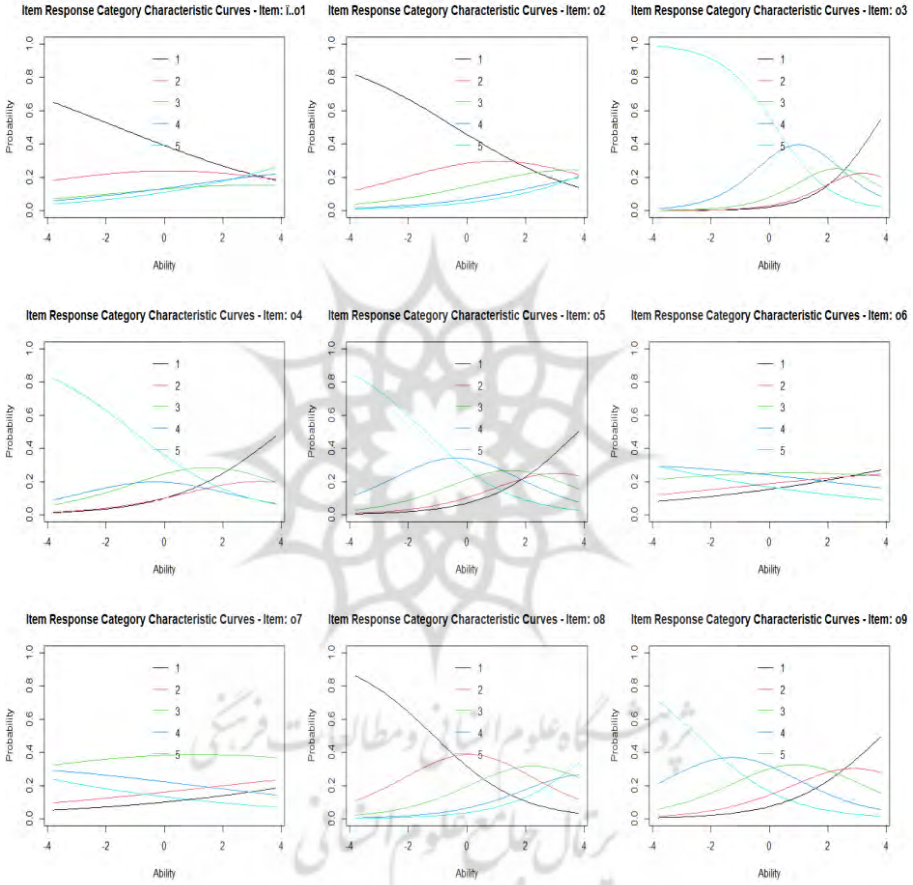
سؤال	ضریب تشخیص	آستانه ۱	آستانه ۲	آستانه ۳	آستانه ٤
۷	-۰/۱۸٦	۱۱/۷۸۸	۵/۵۹۵	-۳/۱۹۳	-۱۰/۰۸۵
۸	۰/٦۸۱	-۱/۱۳۳	۱/۲۷۰	۳/۱۹۹	٤/۷۹۲
۹	-۰/٦۵۹	۳/۸٤۱	۱/۹٤٦	-۱/۱۰۷	-۲/٤٦۱
۱۰	-۱/۰٦٦	۲/۹۸۱	۲/۰۰۲	۰/٦۸۷	-۰/۷۳۲
۱۱	-۰/۹٤۵	۳/۹۸۲	۲/۵۵۵	۱/۰٤۵	-۰/۵۹۹
۱۲	-۰/۲۱٦	۸/٦٤۸	۳/۱۹۲	-۲/۷۳٤	-۸/۳٦۸

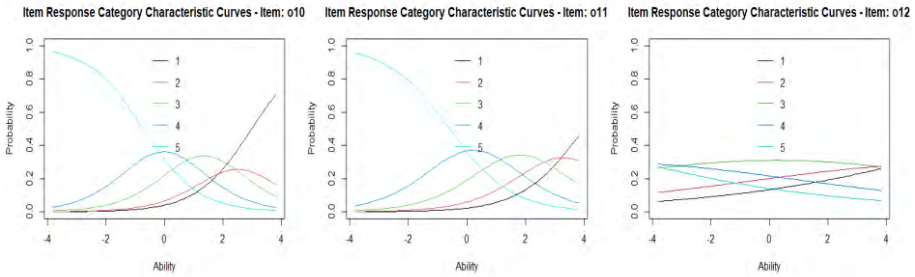
همان‌طور که در جدول بالا مشاهده می‌شود، ضریب تشخیص به‌دست‌آمده از نظریه سؤال پاسخ عکس نتیجه‌ای است که از نظریه کلاسیک به‌دست‌آمده است؛ درواقع در اینجا همه ضرایب تشخیص منفی و تنها ضریب تشخیص مربوط به سؤالات ۱، ۲ و ۸ مثبت شده است. سؤالات ۳ و ۱۰ ضریب تشخیص بالایی به لحاظ بزرگی دارند منتها ضریب تشخیصشان منفی است.

در ستون‌های بعدی جدول میزان آستانه گزارش شده است. تعداد آستانه برای هر سؤال برابر است با تعداد گزینه‌های سؤال منهای یک. با توجه به اینکه در پرسشنامه نئو هر سؤال ۵ گزینه دارد، تعداد آستانه‌ها برابر با ۴ است. میزان آستانه اول برای سؤال یک برابر با ۱/۵۹۲- است، این بدان معناست که شخصی با سطح صفت (گشودگی) برابر با ۱/۵۹۲- احتمال اینکه گزینه ۱ را مقابل گزینه‌های دیگر برگزیند ۵۰ درصد است. آستانه دوم برای این سؤال برابر با ۱/۸۸۶ است، یعنی شخصی با گشودگی ۱/۸۸۶ احتمال اینکه گزینه ۱ و ۲ را در مقابل ۳ گزینه بالاتر برگزیند برابر با ۵۰ درصد است. به لحاظ منطقی باید هر چه به سمت آستانه بالاتر می‌رویم میزان سطح صفت موردمطالعه نیز افزایش یابد؛ ولی همان‌طور که در جدول مشاهده می‌شود این روال تنها برای سؤال ۱، ۲ و ۸ صادق بوده و برای سایر سؤالات عکس این روال صدق می‌کند، برای مثال در سؤال ۳، میزان آستانه اول برابر با ۳/۶۲۹ بوده و هرچه به سمت آستانه‌های بالاتر پیش می‌رویم این میزان کاهش‌یافته و میزان آستانه چهارم به ۰/۲۱ می‌رسد. درواقع سؤالاتی که دارای ضریب تشخیص منفی بودند، میزان آستانه آن‌های به‌جای اینکه سیر صعودی داشته باشند، یک سیر نزولی را نشان دادند. می‌توان برای نمایش آستانه‌های هر سؤال از منحنی ویژه عملیاتی (OCC) که یک مدل لاجیت تجمعی و معادل ICC در مدل‌های دوارزشی است، استفاده کرد که در اینجا به دلیل حجم زیاد اطلاعات گزارش نشده، ولی در پیوست ارائه شده است.

در ادامه منحنی ویژه طبقات پاسخ مربوط به هر سؤال گزارش شده است. منحنی ویژه طبقات که به‌اختصار CCC نامیده می‌شوند نشان می‌دهند که در هر سطح توانایی احتمال انتخاب هر گزینه چقدر است.

شکل ۲. منحنی ویژه طبقات سؤالات گشودگی

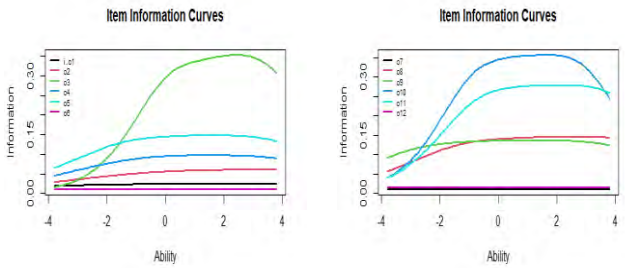




منحنی مشخصات طبقات سؤال ۷ و ۱۲ نشان می‌دهد که در هر سطحی از صفت گشودگی احتمال انتخاب گزینه سوم که بیان‌گر حد وسط است، محتمل‌ترین انتخاب است. در سؤال اول در طیف وسیعی از صفت گزینه ۱ محتمل‌ترین گزینه انتخابی است و نزدیک به سطح صفت برابر با ۴ گزینه آخر احتمال انتخاب بیشتری را دارد؛ درواقع سایر گزینه‌های این سؤال نقش تعیین‌کننده‌ای در تعیین سطح صفت نداشته و وجود آن‌ها در این سؤال مهم نبوده است. در سؤالات ۳، ۴، ۵، ۹، ۱۰ و ۱۱ ترتیب گزینه‌ها عکس شده است، یعنی احتمال انتخاب گزینه‌های با نمره بیشتر در سطح صفت پایین‌تر، بیشتر است؛ این همان نتیجه‌ای است که در مورد ضریب تشخیص و میزان آستانه‌ها نیز شاهد آن بودیم.

درنهایت تابع آگاهی آزمون، محاسبه گردید. آگاهی کل آزمون مبتنی بر این ۱۲ سؤال، برابر با ۰/۶۸ است که در محدوده ۴- تا ۴ سطح صفت که در این تحلیل مدنظر بود، آگاهی برابر با ۰/۳ است، یعنی آگاهی این دامنه از سطح صفت تنها شامل ۴۴/۹ درصد از آگاهی کل است. در شکل زیر منحنی آگاهی هر سؤال که به‌اختصار IIC نامیده می‌شود، نمایش داده شده است. همان‌طور که در تصویر مشاهده می‌شود سؤالات ۳، ۱۰ و ۱۱ در دامنه ۲- تا ۴ بیشترین آگاهی را ارائه می‌کنند، یعنی این سؤالات برای برآورد گشودگی افرادی که نمره آن‌ها در بازه ۲- تا ۴ است، آگاهی‌دهنده‌تر است. سؤالات ۱، ۶، ۷ و ۱۲ نیز در تمام سطوح صفت آگاهی پایین و نزدیک به صفر دارند.

شکل ۳. منحنی آگاهی سؤالات گشودگی



برون‌گرایی: این ویژگی دومین ویژگی شخصیتی موردمطالعه در آزمون شخصیت نئو است که با ۱۲ سؤال اندازه‌گیری می‌شود. اندازه نمونه موردمطالعه در این ویژگی برابر با ۷۸۳ نفر است که میانگین نمرات برون‌گرایی آن‌ها برابر با ۳۲/۱۶ و انحراف استاندارد آن برابر با ۵/۶۶ است. کمترین نمره برون‌گرایی برابر با ۱۳ و بیشترین آن برابر با ۴۵ است. در جدول زیر درصد فراوانی انتخاب هر گزینه توسط افراد، میانگین مربوط به هر سؤال (ضریب مقبولیت)، همبستگی سؤال با نمره کل (ضریب تشخیص بر اساس نظریه کلاسیک آزمون) و آلفا در صورت حذف سؤال گزارش شده است.

جدول ٤. فراوانی درصدی انتخاب هر گزینه، میانگین و همبستگی هر سؤال با نمره کل و ضریب آلفا در صورت حذف سؤال

سؤال	گزینه ۱	گزینه ۲	گزینه ۳	گزینه ٤	گزینه ٥	میانگین	همبستگی	آلفا
۱	۸/۸	۱۷/۲	۱۳/٤	۳٤/٦	۲٥/۹	۲/٥۲	۱/۱۹	۰/٥۱
۲	۹/۱	۱۹/۲	۱۷/۹	۲۹/۹	۲٤	۲/٤۱	۰/۰۳٤	۰/٥٥۲
۳	۷	۹/۸	۱۲/۸	۳۰	٤۰/٤	۲/۸۷	۰/۲٤٥	۰/٤۹٦
٤	٤/٦	٦/۸	۱٥/۱	۳٥/٥	۳۸/۱	۲/۹٦	۰/۳۳٦	۰/٤۷٥
٥	۳/۸	۱۱/۱	۱۹	۳۳/۸	۳۲/۲	۲/۷۹	۰/۲٥۱	۰/٤۹٥
٦	۱۱/٥	۱۷	۱۷/۲	۳٤/۷	۱۹/٥	۲/۳٤	۰/۱۱۲	۰/٥۳۲
۷	۳/۷	٤/٥	۱٥/۲	۳٤/۲	٤۲/٤	۳/۰۷	۰/۳۲٦	۰/٤۷۹
۸	۳/٦	٦/٦	۱۳/٥	۳٦	٤۰/۲	۳/۰۳	۰/۳۹۳	۰/٤٦۲
۹	٤/۷	۸/٤	۱۲/۳	۲٥/۲	٤۹/٤	۳/۰٦	۰/۳٤۳	۰/٤۷۰
۱۰	٦/۸	۱۲/۹	۲۸/٤	۳۰/۸	۲۱/۲	۲/٤۷	۰/۰۸٥	۰/٥۳٥

سؤال	گزینه ۱	گزینه ۲	گزینه ۳	گزینه ٤	گزینه ٥	میانگین	همبستگی	آلفا
۱۱	۲/۳	۷	۲۰/۱	۳٥/٥	۳٥/۱	۲/۹٤	۰/۳۱۹	۰/٤۸۱
۱۲	۲۲/۱	۲۳/۲	۲۷/۲	۱٦/۲	۱۱/۲	۱/۷۱	-۰/۰۰۸	۰/٥٦۳

مطابق با درصد فراوانی‌های انتخاب هر گزینه، مشاهده می‌شود که به‌جز سؤال ۱۲ در مورد سایر سؤالات اقبال نسبت به گزینه‌های پایانی بیشتر از گزینه‌های ابتدایی است؛ یعنی بیشتر افراد تمایل به برون‌گرایی داشته‌اند. در سؤالات ۳، ۴، ۷، ۸، ۹ و ۱۱ بیش از ۷۰ درصد افراد دو گزینه انتهایی را انتخاب کرده‌اند. بیشترین ضریب مقبولیت برای سؤال ۷ و برابر با ۳/۰۷ و کمترین آن مربوط به سؤال ۱۲ و برابر با ۱/۷۱ است. با توجه به محتوای این دو سؤال و در نظر گرفتن این نکته که سؤالات معکوس قبل از انجام هر تحلیلی در جهت سازه برون‌گرایی نمره‌گذاری شده‌اند، می‌توان گفت که بیشتر افراد خود را فردی سرشار از انرژی دانسته، ولی ترجیح می‌دهند به‌تنهایی کار کنند تا اینکه دیگران را راهبری کنند.

سؤال ۱۲ دارای ضریب تشخیص منفی بوده و با سایر سؤالات مربوط به برون‌گرایی رابطه منفی دارد. از طرف دیگر با توجه به ضریب آلفای کرونباخ مربوط به ویژگی برون‌گرایی که برابر با ۰/۵۲۷ است، در صورت حذف سؤالات ۲، ۶، ۱۰ و ۱۲ ضریب آلفا افزایش می‌یابد. این نتیجه نشان می‌دهد که این ۴ سؤال تهدیدی برای همسانی درونی سؤالات این مجموعه هستند. در ادامه تحلیل مبتنی بر نظریه سؤال پاسخ که با مدل "GRM" انجام‌شده گزارش شده است. در جدول زیر ضریب تشخیص و آستانه‌های مربوط به هر سؤال گزارش شده است.

جدول ۵. ضریب تشخیص و آستانه‌های مربوط به هر سؤال

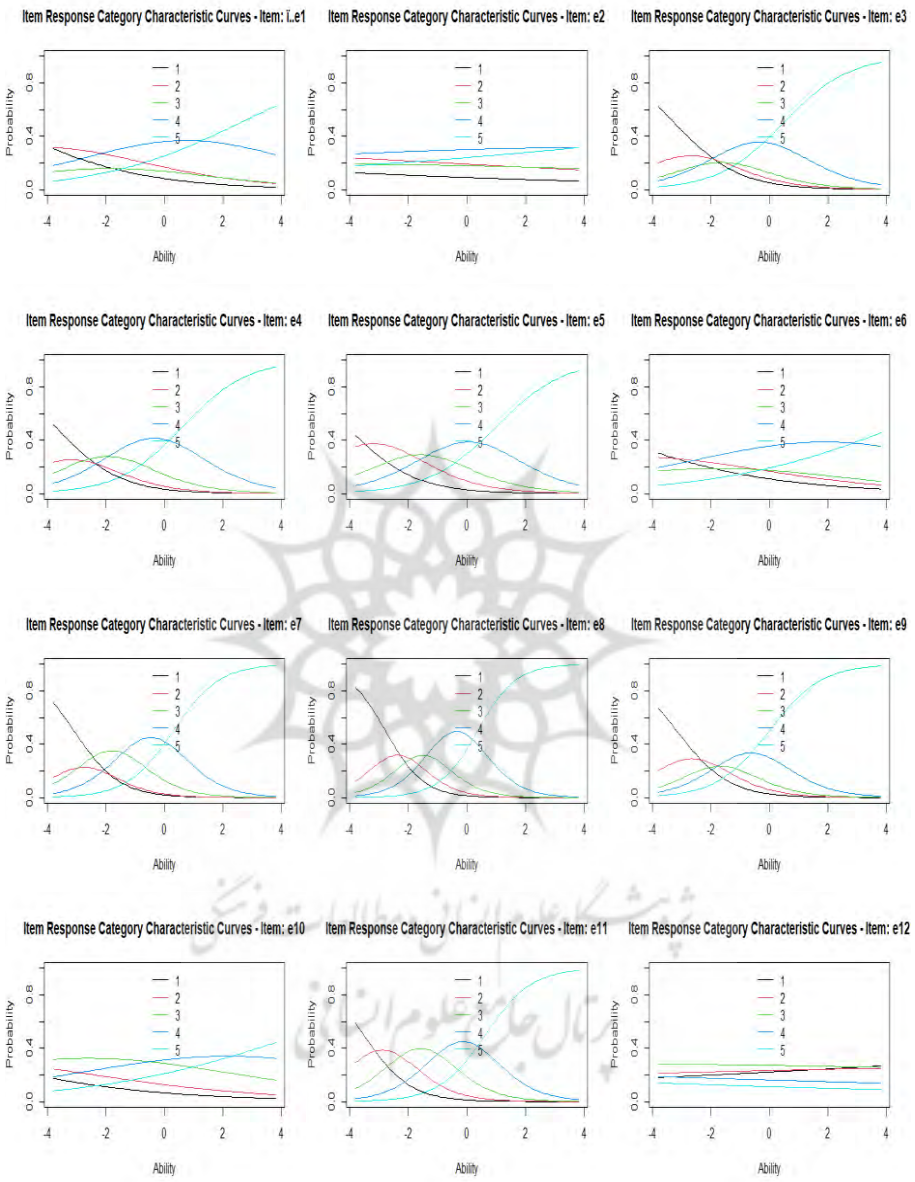
سؤال	ضریب تشخیص	آستانه ۱	آستانه ۲	آستانه ۳	آستانه ٤
۱	۰/٤۲۰	-٥/۷٤۹	-۲/٦۱۳	-۱/۰۹٦	۲/٥۸۱
۲	۰/۰۹۹	-۲۳/۳۸۱	-۹/٥۰۳	-۱/٦۲٤	۱۱/٦٥۹
۳	۰/۸۹۹	-۳/۲٤۹	-۲/۰۹٤	-۱/۱۷٦	۰/٤۸۲
٤	۰/۹۰۹	-۳/۷۲۳	-۲/٥۸۱	-۱/۳۲٤	۰/٦۱۸
٥	۰/۸٦۷	-٤/۰۸٥	-۲/۲۷۳	-۰/۸۸۹	۰/۹۹۱
٦	۰/۳۳۰	-٦/۳۲۲	-۲/۲۹۱	-۰/٥۸٥	٤/۳٥٥
۷	۱/۲۷٥	-۳/۰۹٤	-۲/۳٦۹	-۱/۲۱۹	۰/۳۰۲

سؤال	ضریب تشخیص	آستانه ۱	آستانه ۲	آستانه ۳	آستانه ٤
۸	۱/۵۳۰	-۲/۷۸۲	-۱/۹۱۸	-۱/۰٦۲	۰/۳۵۹
۹	۱/۱۱۹	-۳/۱۷۲	-۲/۱۰۲	-۱/۲٤۸	-۰/۰۰٦
۱۰	۰/۲۸۹	-۹/۱۸۸	-٤/۹۵۰	-۰/۲۹٦	٤/۵۹٦
۱۱	۱/۲٦۵	-۳/۵۲۱	-۲/۲۳۷	-۰/۹۰٦	۰/٦۲۲
۱۲	-۰/۰٦٦	۱۸/۹۷۲	۲/۷۷۲	۱٤/٦۹۹	-۳۱/۱۹۳

با بررسی ستون مربوط به ضرایب تشخیص، مشاهده می‌شود که ضریب تشخیص سؤال ۱۲ منفی شده است، یعنی سؤال ۱۲ عکس سایر سؤالات عمل کرده و تشخیص اشتباه در مورد برآورد برون‌گرایی افراد می‌دهد و این با نتیجه به‌دست‌آمده از نظریه کلاسیک آزمون همخوانی دارد. سؤالات ۸، ۷ و ۱۱ به ترتیب دارای بالاترین ضریب تشخیص بین سؤالات هستند.

در ستون‌های بعدی جدول میزان آستانه گزارش شده است. در تمامی سؤالات به‌جز سؤال ۱۲ هرچه به سمت آستانه بالاتر می‌رویم، میزان آستانه افزایش می‌یابد؛ ولی سؤال ۱۲ به دلیل عملکرد معکوس در این مجموعه، میزان آستانه چهارم آن از همه آستانه‌ها کوچک‌تر شده است. منحنی ویژه عملیاتی منطبق بر آستانه‌ها در پیوست ارائه شده است. در ادامه منحنی ویژه طبقات پاسخ مربوط به هر سؤال گزارش شده است.

شکل ٤. منحنی ویژه طبقات سؤالات برون‌گرایی

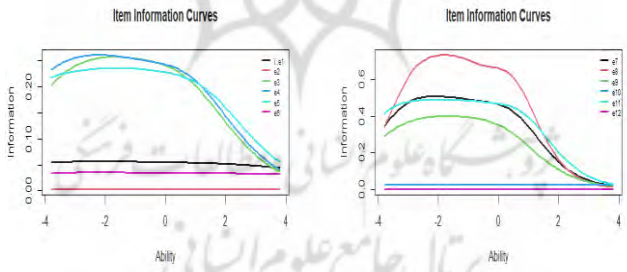


منحنی ویژه طبقات برای سؤال ۲ نشان می‌دهد که در سطح گسترده‌ای از صفت مکنون (برون‌گرایی) احتمال انتخاب گزینه ۴ و برای سؤال ۱۲ احتمال انتخاب گزینه ۳ از سایر گزینه‌ها بیشتر بوده است؛ یعنی سایر گزینه‌های این دو سؤال نقش تعیین‌کننده‌ای در تعیین

سطح صفت نداشته و وجود آن‌ها در این سؤال مهم نبوده است. منحنی ویژه طبقات سؤالات ۵، ۸ و ۱۱ نشان می‌دهد که در این ۳ سؤال تمامی گزینه‌ها مؤثر بوده و هرکدام از گزینه‌ها در یک بازه‌ای از سطح صفت برآورد شده بیشترین احتمال انتخاب را نسبت به سایر گزینه‌ها دارند، ضمن اینکه توالی بین گزینه‌ها به‌خوبی حفظ شده است، یعنی گزینه با امتیاز کمتر برای سطح صفت پایین‌تر محتمل‌تر است. همچنین گزینه ۳ برای سؤال ۱ و گزینه‌های ۱ و ۲ برای سؤال ۱۰ در هیچ سطحی از برون‌گرایی گزینه محتمل نبوده‌اند و این سؤالات می‌توانستند با تعداد گزینه محدودتری نیز ارائه شوند.

درنهایت تابع آگاهی آزمون، محاسبه گردید. آگاهی کل آزمون مبتنی بر این ۱۲ سؤال، برابر با ۰/۸۳ است که در محدوده ۴- تا ۴ سطح صفت که در این تحلیل مدنظر بود، آگاهی برابر با ۰/۴۳ است، یعنی آگاهی این دامنه از سطح صفت تنها شامل ۵۲/۱۱ درصد از آگاهی کل است. در شکل زیر منحنی آگاهی هر سؤال نمایش داده شده است. همان‌طور که در تصویر مشاهده می‌شود سؤال ۸ بیشترین آگاهی را در بازه ۴- تا ۲ نشان داده است و بعدازآن سؤالات ۷، ۱۱ و ۹ به ترتیب آگاهی بیشتری نشان دادند. سؤالات ۲، ۱۰ و ۱۲ نیز در تمام سطوح صفت آگاهی پایین و نزدیک به صفر دارند.

شکل ۵. منحنی آگاهی سؤالات برون‌گرایی



مسئولیت‌پذیری: این ویژگی پنجمین ویژگی شخصیتی موردمطالعه در آزمون شخصیت نئو است که با ۱۲ سؤال اندازه‌گیری می‌شود. اندازه نمونه موردمطالعه در این ویژگی برابر با ۷۸۴ نفر است که میانگین نمرات مسئولیت‌پذیری آن‌ها برابر با ۳۶/۱۱ و انحراف استاندارد آن برابر با ۷/۰۱ است. کمترین نمره مسئولیت‌پذیری برابر با ۱۳ و بیشترین آن برابر با ۴۸ است. در جدول زیر درصد فراوانی انتخاب هر گزینه توسط افراد، میانگین مربوط به هر

سؤال (ضریب مقبولیت)، همبستگی سؤال با نمره کل (ضریب تشخیص بر اساس نظریه کلاسیک آزمون) و آلفا در صورت حذف سؤال گزارش شده است.

جدول ٦. فراوانی درصدی انتخاب هر گزینه، میانگین و همبستگی هر سؤال با نمره کل و ضریب آلفا در صورت حذف سؤال

سؤال	گزینه ۱	گزینه ۲	گزینه ۳	گزینه ٤	گزینه ۵	میانگین	همبستگی	آلفا
۱	۳/۳	٤	۸/۸	۳۱/۹	۵۲	۳/۲۵	۰/۳۵۳	۰/۷٤۷
۲	٤	۸/۹	۱۸/۱	۳۸	۳۱	۲/۸۳	۰/۳۵۳	۰/۷٤۷
۳	٦/٤	۹/۱	۱۲	۳۵/۳	۳۷/۲	۲/۸۸	۰/۳۸٤	۰/۷٤٤
٤	۲/۷	۲/٦	۷/۹	۲٦/۷	٦۰/۲	۳/۳۹	۰/٤۷۰	۰/۷۳٦
۵	٤/۷	٦/۱	۱۵/۳	۲۸/۸	٤۵	۳/۰۳	۰/٤۸۲	۰/۷۳۲
٦	۵/۲	۸/۵	۱٦/۱	۳۱/۹	۳۸/۳	۲/۸۹	۰/۳۷٤	۰/۷٤۵
۷	۳/۲	۵/۱	۷	۲٤/۵	٦۰/۲	۳/۳۳	۰/٤۵۵	۰/۷۳٦
۸	٤/٦	٦/۹	۲۰/۸	۳۱/۳	۳٦/۵	۲/۸۸	۰/۳۷۰	۰/۷٤۵
۹	۵/۷	۱٦/۳	۲۵/٦	۲٦	۲٦/۳	۲/۵۱	۰/۳۳٤	۰/۷۵۰
۱۰	۳/۸	۸/۲	۱۸/۵	۳۳/۸	۳۵/۷	۲/۸۹	۰/٤۰۲	۰/۷٤۱
۱۱	۷/۵	۸/۷	۱٤/۵	۲۲/۳	٤٦/۹	۲/۹۲	۰/۳٦۷	۰/۷٤٦
۱۲	۵	٤/۲	۱۱/۹	۱۵/۲	٦۳/۸	۳/۲۹	۰/٤۰۷	۰/۷٤۱

مطابق با درصد فراوانی‌های انتخاب هر گزینه، مشاهده می‌شود که اقبال شرکت‌کنندگان به دو گزینه انتهایی بسیار زیاد است؛ یعنی بیشتر افراد خودشان را فردی مسئولیت‌پذیر می‌دانند. در سؤالات ۱، ۴ و ۷ بیش از ۸۰ درصد افراد دو گزینه انتهایی را انتخاب کرده‌اند. بیشترین ضریب مقبولیت برای سؤال ۴ و برابر با ۳/۳۹ است، یعنی بیشتر افراد ادعا کرده‌اند که سعی می‌کنند کارهایشان را با احساس مسئولیت انجام دهند.

ضریب تشخیص تمامی سؤالات در این بخش مثبت بوده و بین سؤالات و نمره کل همبستگی مطلوبی وجود دارد. از طرف دیگر با توجه به ضریب آلفای کرونباخ مربوط به ویژگی مسئولیت‌پذیری که برابر با ۰/۷۵۹ است، هیچ سؤالی وجود ندارد که در صورت حذف آن ضریب آلفا افزایش یابد. در ادامه تحلیل مبتنی بر نظریه سؤال پاسخ که با مدل "GRM" انجام‌شده گزارش شده است. در جدول زیر ضریب تشخیص و آستانه‌های مربوط به هر سؤال گزارش شده است.

جدول ۷. ضریب تشخیص و آستانه‌های مربوط به هر سؤال

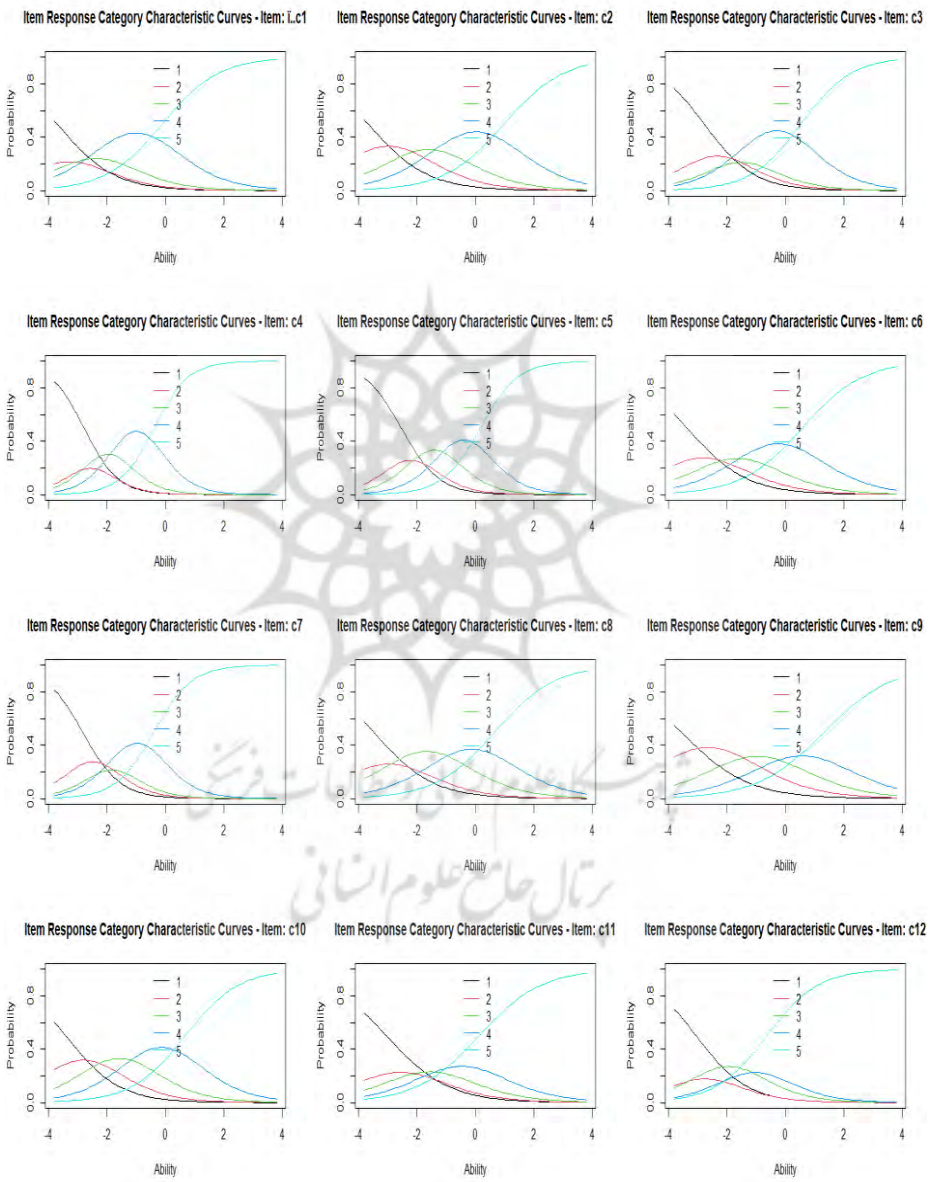
سؤال	ضریب تشخیص	آستانه ۱	آستانه ۲	آستانه ۳	آستانه ٤
۱	۱/۰۳۱	-۳/۷۰۱	-۲/۸۵۳	-۱/۹۰۱	-۰/۱۰۹
۲	۰/۹۷٤	-۳/٦۷۳	-۲/۲٥۷	-۰/۹٥٦	۰/۹۸۳
۳	۱/۱٥۸	-۲/۷۸٥	-۱/۸۷۲	-۱/۱۳۳	۰/٥۳۰
٤	۱/۷۰۷	-۲/۷۸۹	-۲/۳۲۲	-۱/٥۹۱	-۰/۳۸٤
٥	۱/٥٤۱	-۲/٥٤۲	-۱/۸٦۱	-۰/۹٦۳	۰/۱٦۷
٦	۰/۹۷۰	-۳/۳۷۳	-۲/۲۱۳	-۱/۰۷۱	۰/٥۸۱
۷	۱/٥٤۰	-۲/۸٤۰	-۲/۱۰٤	-۱/٥٤۰	-۰/۳۹٤
۸	۰/۹۸۳	-۳/٤۸٦	-۲/٤۱۰	-۰/۹۰۹	۰/٦٦۷
۹	۰/۸٦٦	-۳/٥۷۱	-۱/۷۰۹	-۰/۱۹۳	۱/۳۳۹
۱۰	۱/۰۹۰	-۳/٤۰٥	-۲/۱۹٦	-۰/۹۳۸	۰/٦۷۳
۱۱	۰/۹٤۱	-۳/۰٤۱	-۲/۰٦۱	-۱/۰٦٤	۰/۱۲۲
۱۲	۱/۱۳۷	-۳/۰٥۳	-۲/٤۲۳	-۱/٤٤٦	-۰/٦٤۰

با بررسی ستون مربوط به ضرایب تشخیص، مشاهده می‌شود که مطابق با نتیجه نظریه کلاسیک آزمون تمامی سؤالات دارای ضریب تشخیص مثبت هستند. سؤال ۴ دارای بالاترین ضریب تشخیص و سؤال ۹ دارای کمترین ضریب تشخیص در بین سؤالات هستند.

در ستون‌های بعدی جدول، میزان آستانه گزارش شده است. سؤالات ۱، ۴، ۷ و ۱۲ هر چهار آستانه‌شان منفی شده است، یعنی حتی زمانی که فردی مسئولیت‌پذیری پایینی هم داشته باشد، احتمال انتخاب گزینه‌های انتهایی و با نمره بالا را در این سؤالات دارد و این سؤالات برای تفکیک افراد با مسئولیت بالا از افراد با مسئولیت متوسط چندان کارا نبوده و دارای اثر سقف هستند. ممکن است این ابهام ایجاد شود که چطور سؤال ۴ دارای ضریب تشخیص بالایی است؛ ولی در اینجا مطرح‌شده که در تفکیک بین افراد با سطوح صفت بالا ناتوان است؛ این مسئله را می‌شود این‌گونه توجیه کرد که گزینه‌های سؤال ۴ در همان محدوده صفتی که بیشترین احتمال انتخاب را دارند به‌خوبی عمل می‌کنند، برای مثال آستانه چهارم که مرز بین انتخاب گزینه آخر و گزینه‌های قبل از آن است در سطح صفت برابر با ۰/۳۸۴- دارای قدرت تشخیص بالایی است. با مراجعه به منحنی ویژه عملیاتی که در پیوست ارائه شده است، مشاهده می‌شود که منحنی‌های لوجیت تجمعی مربوط به سؤال ۴ دارای بیشترین

شیب نسبت به سایر گزینه‌ها هستند و همین امر باعث ضریب تشخیص بالای این سؤال شده است. در ادامه منحنی ویژه طبقات پاسخ مربوط به هر سؤال گزارش شده است.

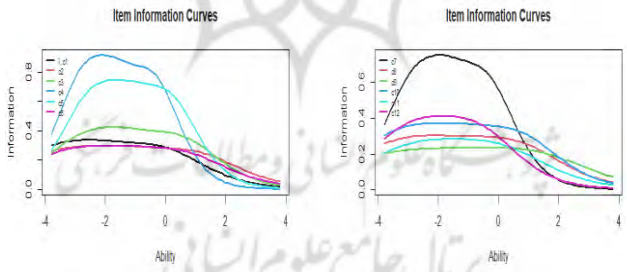
شکل ٦. منحنی ویژه طبقات سؤالات مسئولیت‌پذیری



منحنی ویژه طبقات سؤالات ۲، ۹ و ۱۰ نشان می‌دهد که در این ۳ سؤال تمامی گزینه‌ها مؤثر بوده و هرکدام از گزینه‌ها در یک بازه‌ای از سطح صفت برآورد شده بیشترین احتمال انتخاب را نسبت به سایر گزینه‌ها دارند، ضمن اینکه توالی بین گزینه‌ها به‌خوبی حفظ شده است. گزینه ۲ در سؤالات ۴، ۵، ۶، ۷، ۸ و ۱۱ برای هیچ میزانی از مسئولیت‌پذیری گزینه محتمل نبوده و این نتیجه نشان می‌دهد که گزینه ۲ در برآورد صفت مسئولیت‌پذیری منفعل عمل کرده و تأثیر چندانی نداشته است.

درنهایت تابع آگاهی آزمون، محاسبه گردید. آگاهی کل آزمون برابر با ۴/۳۷ است که در محدوده ۴- تا ۴ سطح صفت که در این تحلیل مدنظر بود، آگاهی برابر با ۳/۶۱ است، یعنی آگاهی این دامنه از سطح صفت شامل ۸۲/۶۵ درصد از آگاهی کل است. در شکل زیر منحنی آگاهی هر سؤال نمایش داده شده است. همان‌طور که در تصویر مشاهده می‌شود در بازه ۴- تا صفر از صفت مکنون (مسئولیت‌پذیری) سؤال ۴ بیشترین آگاهی را نشان داده است و بعدازآن سؤال ۷ در بازه ۴- تا ۲- و سؤال ۵ در بازه ۲- تا صفر آگاهی بیشتری نشان دادند. هیچ سؤالی در این مجموعه، آگاهی نزدیک به صفر نداشت؛ ولی آگاهی همه سؤالات بعد از سطح صفت ۲ به صفر میل می‌کند.

شکل ۷. منحنی آگاهی سؤالات مسئولیت‌پذیری



بحث و نتیجه‌گیری

بر اساس اهداف، در این پژوهش تلاش شد تا نتایج تحلیل سؤالات بر اساس دو نظریه کلاسیک و سؤال پاسخ در کنار یکدیگر قرار گرفته و تفاوت و تشابه آن‌ها مشخص شود. به همین منظور سه دسته سؤال با ضرایب همسانی متفاوت انتخاب و مورد تجزیه‌وتحلیل قرار گرفتند. نتایج ضریب آلفای کرونباخ که منطبق بر نظریه کلاسیک است با نتایج مربوط به

تابع آگاهی که مبتنی بر نظریه سؤال پاسخ است، همخوانی نشان داد؛ یعنی همان‌طور که سؤالات گشودگی کمترین ضریب آلفا را نشان داد، دارای کمترین آگاهی هم بود و سؤالات مسئولیت‌پذیری نیز دارای بالاترین میزان آلفا و آگاهی در بین این سه مجموعه سؤال بود و در هر دو مورد سؤالات برون‌گرایی حد وسط را نشان دادند. البته از طریق بررسی تابع آگاهی در نظریه سؤال پاسخ نتایج بیشتری قابل حصول است؛ مثلاً در نظریه سؤال پاسخ این امکان وجود دارد که علاوه بر تابع آگاهی کلی برای هر سؤال به‌طور جداگانه تابع آگاهی به دست آورده و همچنین مشخص کرد که در کدام سطوح توانایی هر سؤال آگاهی‌دهنده‌تر است؛ ولی در نظریه کلاسیک چنین امکانی وجود ندارد. علاوه بر این در بررسی توابع آگاهی مربوط به هر سؤال مشخص شد که سؤالات ۲، ۱۰ و ۱۲ در ویژگی برون‌گرایی و سؤالات ۱، ۶، ۷، و ۱۲ در ویژگی گشودگی در تمامی سطوح، آگاهی نزدیک به صفر دارند و حذف این سؤالات از مجموعه سؤالات آسیبی به آگاهی آزمون نزده بلکه طول آزمون را کوتاه‌تر کرده و باعث صرفه‌جویی می‌شود؛ درصورتی‌که در تحلیل سؤالات با رویکرد کلاسیک، سؤالات ۲ و ۱۰ برون‌گرایی و سؤالات ۷ و ۱۲ گشودگی مشکل‌دار تشخیص داده نشدند. این یافته‌ها می‌تواند مهر تأییدی بر عقیده Baker and Kim (2004) باشد که نظریه سؤال پاسخ را بسط و توسعه‌ای از رویکرد کلاسیک با ریشه‌های ریاضی می‌دانست، درواقع IRT چیزی جدای از نظریه کلاسیک نبوده بلکه توسعه‌یافته و تکمیل شده آن است.

با اینکه در ضریب آلفا و آگاهی تناقضی مشاهده نشد؛ اما در بررسی ضرایب تشخیص از طریق دو رویکرد، تناقض آشکاری در تحلیل سؤالات گشودگی پدیدار شد. دقیقاً سه سؤالی که در رویکرد کلاسیک ضریب تشخیص منفی نشان دادند در نظریه سؤال پاسخ ضریب تشخیص مثبت داشتند و بالعکس که همین امر موجب شد تا منحنی ویژه طبقات سؤالات در ویژگی گشودگی دچار وارونگی شده و انتخاب گزینه با امتیاز بیشتر در سطح صفت پایین‌تر برای بیشتر سؤالات احتمال بالایی نشان دهد. با بررسی‌های انجام‌شده و با انجام تحلیل‌های مختلف مشخص شد که علت رخ دادن چنین اتفاقی، این بود که اولین سؤال مجموعه دارای همبستگی معکوس با نمره کل آزمون بود (نتیجه حاصل از تحلیل کلاسیک آزمون) و در تحلیل مبتنی بر IRT همان سؤال اول، مبنا قرار گرفته و صفت مکنون بر اساس سؤال اول برآورد شده و همین امر سبب منفی‌شدن ضرایب تشخیص ۹ سؤال دیگر

که با سؤال اول همبستگی منفی داشتند، شد و زمانی که جای سؤال اول و سؤال سوم در فایل داده‌ها تغییر کرد، علامت ضرایب تشخیص برآورد شده با IRT مشابه با علامت ضرایب تشخیص در نظریه کلاسیک آزمون شد. با بررسی‌های انجام‌شده توسط پژوهشگر در پژوهش یا کتابی به این نکته اشاره نشده؛ درصورتی‌که این می‌تواند یکی از ضعف‌های بزرگ مدل مدرج پاسخ بوده و اگر تحلیل کلاسیک آزمون انجام نشده بود، پژوهشگر را به اشتباه و گمراهی می‌کشاند.

با جمع‌بندی این یافته‌ها به این نتیجه می‌رسیم که در صورت وجود سؤالات ضعیف امکان به اشتباه افتادن و نتیجه‌گیری غلط از تحلیل‌های حاصل از نظریه سؤال پاسخ افزایش می‌یابد. همان‌طور که مشاهده شد در مورد سؤالات گشودگی که دارای ضریب آلفا بسیاری پایینی بودند و همبستگی بین سؤالات پایین و گاهی منفی بود، این اتفاق بیشتر نمایان شد. همچنین میزان آستانه در سؤالات این مجموعه به اعداد نامعقولی مثل ۱۱ و ۱۰ نیز می‌رسید؛ درصورتی‌که بر اساس نظر Baker and Kim (2004) نمرات صفت پنهان و پارامترهای آستانه معمولاً از ۲- تا ۲ متغیر است، حتی آستانه بین ۳- و ۳ نیز غیرمعقول نیست؛ ولی مقادیر آستانه خارج از این محدوده غیرمعمول هستند و به‌طور بالقوه نشانه‌ای از موارد مشکل‌ساز یا دسته‌های پاسخ کمتر مفید هستند؛ بنابراین تلاش برای استفاده از مقیاسی با ویژگی‌های روان‌سنجی ضعیف (مانند دقت ضعیف در سراسر یا در مکان‌های خاص در امتداد زنجیره صفت پنهان موردنظر) می‌تواند استنتاج‌های تفسیر نمره بالقوه را به خطر بیندازد و اگر به‌عنوان متغیر برونداد در یک تحلیل آماری استفاده شود، می‌تواند منجر به کاهش روایی نتایج آماری شود (Kang & Waller, 2005). نتیجه پژوهش زمانپور و همکاران (۱۳۹۸) نیز نشان داد که اگر مفروضه تک‌بعدی بودن به‌طور سخت‌گیرانه‌ای رعایت شود، گویه‌هایی که مناسب تشخیص داده می‌شوند در هر دو تحلیل یکسان خواهد بود؛ بنابراین اگر مفروضه‌ها به‌درستی رعایت شده و سؤالات دارای ویژگی‌های روان‌سنجی مناسبی باشند، تفاوت زیادی بین دو تحلیل مشاهده نمی‌شود؛ اگرچه تحلیل‌های IRT امکان ارائه اطلاعات جامع‌تری را دارند. برای مثال در تحلیل و واکاوی سؤالات مسئولیت‌پذیری، برتری مدل‌های IRT مشخص شد؛ زیرا اطلاعاتی که در مورد تابع آگاهی این سؤالات مشاهده شد از طریق رویکرد کلاسیک دست‌یافتنی نبود. در تحلیل سؤالات مسئولیت‌پذیری با اینکه در رویکرد کلاسیک همه

شرایط مناسب نشان داده می‌شد و ردی از مشکل نبود با تحلیل IRT مشخص شد که این سؤالات برای تشخیص افراد با مسئولیت بالاتر از ۲ ناتوان بوده و آگاهی‌دهنده نیستند.

توصیه می‌شود که قبل از انجام هر تجزیه‌وتحلیل IRT، بررسی شود که آیا در هر دسته فراوانی به اندازه بوده و همه گزینه‌ها توسط افراد گروه نمونه به حد کفایت انتخاب شده‌اند؛ درواقع در این پژوهش مشاهده شد که در سؤالات مشکل‌دار بیشتر افراد گزینه‌های خاصی را انتخاب کرده و همین امر باعث نامعقول شدن منحنی ویژه طبقات شد. اگرچه هیچ دستورالعمل سخت و صریحی برای ملاک کافی بودن در نظر گرفته نشده، پاسخ‌های بیشتر در هر دسته به افزایش دقت برآورد پارامتر سؤال و ارزیابی سودمندی دسته‌های پاسخ کمک می‌کند. اگر از اعداد کافی در دسته‌های پاسخ استفاده نمی‌شود، ممکن است لازم باشد دسته‌های پاسخ با یکدیگر تجمیع و ادغام شوند تا یک سیستم دسته‌بندی پاسخ کاهش‌یافته تولید شود. انجام این کار باعث بهبود دقت و پایداری در برآورد پارامتر سؤال می‌شود (تولند، ۲۰۱۴). همچنین مشابه با این پژوهش بهتر است که در بررسی سؤالات هر آزمونی از روش‌های هر دو نظریه استفاده شده تا ضعف‌هایی که هر نظریه دارد و در این پژوهش به‌صورت عملی بررسی و گزارش شد توسط رویکرد دیگر پوشش داده شد و بهترین ارزیابی از آزمون موردبررسی به عمل آید. همچنین پیشنهاد می‌شود که در مطالعات آتی سایر مدل‌های نظریه سؤال پاسخ با تحلیل‌ها نظریه کلاسیک آزمون مقایسه شود تا نقاط ضعف و قوت هر رویکرد به‌صورت عملی مشاهده و موردبحث قرار گیرد.

تعارض منافع

تعارض منافع ندارم.

منابع

تیلور، کاترین، اس. (۱۳۹۸). *روایی و رواسازی*. (ترجمه یونسی، جلیل). تهران: دانشگاه علامه طباطبائی. https://book.atu.ac.ir/

زمانپور، عنایت اله، یونسی، جلیل، رستگار آگاه، مصطفی، و مهرابی، مهسا. (۱۳۹۸). تحلیل داده های نگرش‌سنجی: تفاوت نظریه کلاسیک و سؤال ـ پاسخ. *پژوهش‌های ارتباطی*. ۲۶ (۲)، ۱۶۵-۱۳۹. https://doi.org/10.22082/cr.2019.113772.1917

References

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques*. CRC press. https://doi.org/10.1201/9781482276725

Baker, F. B. (2001). *The basics of item response theory*. For full text: http://ericae.net/irt/baker

Courville, T. G. (2004). *An empirical comparison of item response theory and classical test theory item/person statistics*. Texas A&M University. https://www.proquest.com/openview/dc9dbe1471e8f2828066045a2396ea24/1?pq-origsite=gscholar&cbl=18750&diss=y

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications. https://cehs.unl.edu/EdPsych/RJSite/de_Ayala_Appendices_2ndEd.pdf

Embretson, S. E., & Reise, P. E. (2000). *Item response theory for psychologists*. Psychology Press. https://doi.org/10.4324/9781410605269

Hambleton, R. K., & Jones, R. W. (1993). An NCME Instructional Module on Educational Measurement: issues and practice. *Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development, 12*(3), 38-47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1985). *Principles and applications of item response theory*.

Hambleton, R. K., & Swaminathan, H. (2013). *Item response theory: Principles and applications*. Springer Science & Business Media. https://link.springer.com/book/10.1007/978-94-017-1988-9

Jeong, H. J., & Lee, W. C. (2016). Item response theory-based evaluation of psychometric properties of the safety attitudes questionnaire-Korean version (SAQ-K). *Biometrics & Biostatistics International Journal, 3*(5), 1-15. http://dx.doi.org/10.15406/bbij.2015.02.00020

Kang, S. M., & Waller, N. G. (2005). Moderated multiple regression, spurious interaction effects, and IRT. *Applied Psychological Measurement, 29*(2), 87-105. https://doi.org/10.1177/0146621604272737

Lord, F.M., Novick, M.R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Addison-Wesley. https://psycnet.apa.org/record/1968-35040-000

Ostini, R., & Nering, M. L. (2006). *Polytomous item response theory models* (No. 144). Sage. https://scholar.archive.org/work/dbmqoxx2jvf6bl3ze75bobgxdq/access/wayback/http://hbanaszak.mjr.uw.edu.pl/TempTxt/ebooksclub.org__Polytomous_Item_Response_Theory_Models__Quantitative_Applications_in_the_Social_Sciences_.pdf

Reeve, B. B., & Fayers, P. (2005). Applying item response theory modeling for evaluating questionnaire item and scale properties. *Assessing quality of life in clinical trials: methods of practice, 2*, 55-73. http://dx.doi.org/10.1093/oso/9780198527695.003.0005

Sorrel, M. A., Olea, J., Abad, F. J., de la Torre, J., Aguado, D., & Lievens, F. (2016). Validity and reliability of situational judgement test scores: A new approach based on cognitive diagnosis models. *Organizational Research Methods, 19*(3), 506-532. https://doi.org/10.1177/1094428116630065

Toland, M. D. (2014). Practical guide to conducting an item response theory analysis. *The Journal of Early Adolescence, 34*(1), 120-151. http://dx.doi.org/10.1177/0272431613511332

Taylor, C. S. (2013). *Validity and validation*. Oxford University Press, USA. [In Persian] https://doi.org/10.1093/acprof:osobl/9780199791040.001.0001

Zamanpour, E., Younesi, J., Rastegar Agah, M., & Mehrabi, M. (2019). Attitudinal Analysis: The Difference Between Classical Test Theory and Item Response Theory. *Communication Research, 26*(2), 139-165. [In Persian] https://doi.org/10.22082/cr.2019.113772.1917

پیوست

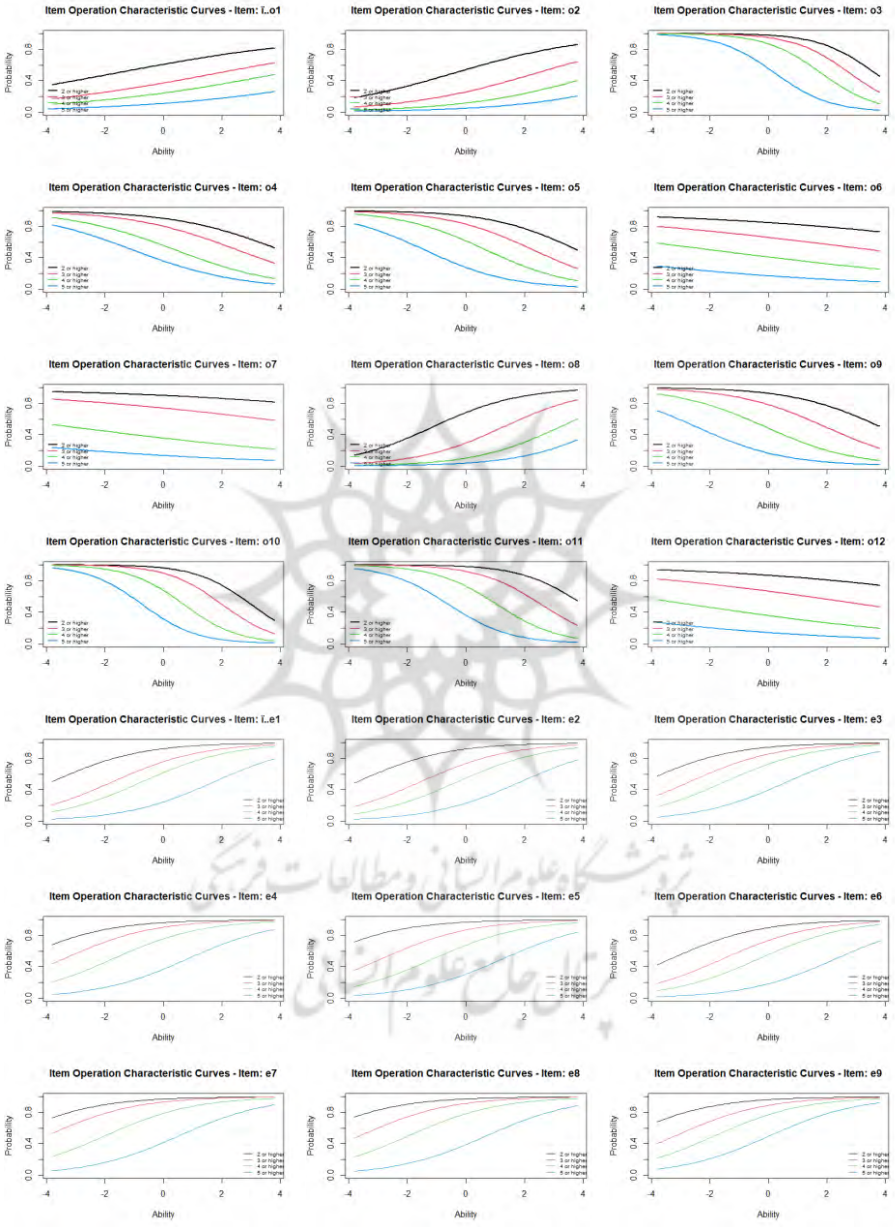
در جدول زیر سؤالات مربوط به هر ویژگی همراه با شماره‌ای که در این تحلیل برای آن استفاده شده به‌علاوه نمادی که در نمودارها برای آن در نظر گرفته شده، آورده شده است.

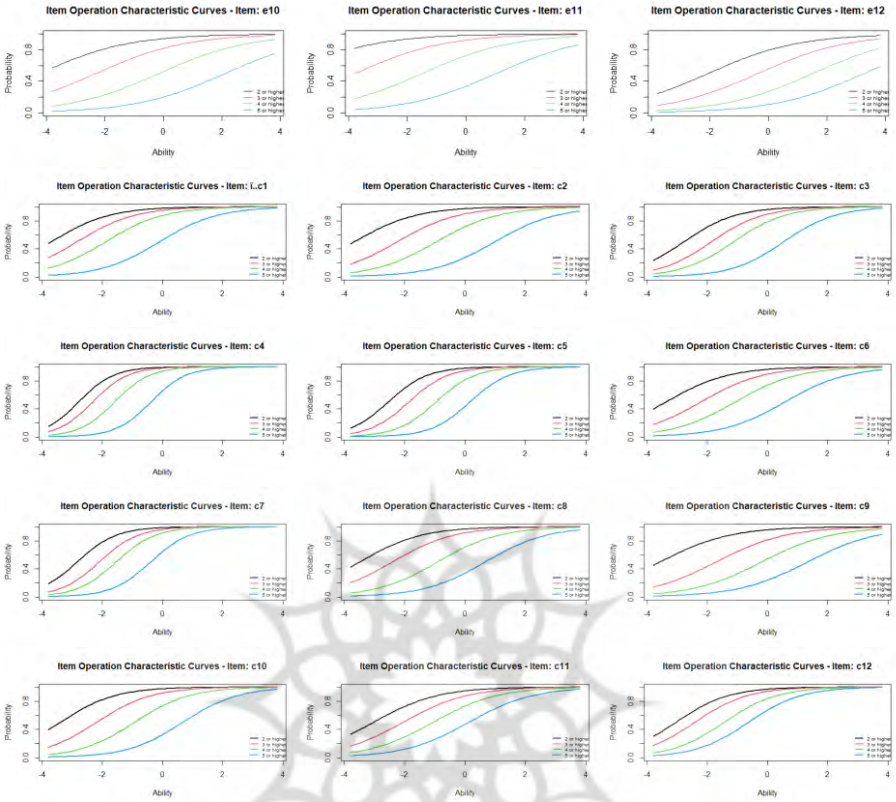
سؤالات گشودگی					
صورت سؤال	شماره نماد		صورت سؤال	شماره نماد	
دوست ندارم وقتم را با خیال‌پردازی تلف کنم.	۱	O۱	به احساسات و عواطفی که محیط‌های متفاوت به وجود می‌آورند، به‌ندرت توجه می‌کنم.	۷	O۷
هنگامی‌که راه درست‌کاری را پیدا کنم، آن روش را همیشه در آن مورد تکرار می‌کنم.	۲	O۲	معتقدم که هنگام تصمیم‌گیری درباره مسائل اخلاقی باید از مراجع مذهبی پیروی کنیم.	۸	O۸
نقش‌های موجود در پدیده‌های هنری و طبیعت من را مبهوت می‌کند.	۳	O۳	بعضی‌اوقات وقتی شعری را می‌خوانم یا یک کار هنری را تماشا می‌کنم، یک احساس لرزش و یک تکان هیجانی را حس می‌کنم.	۹	O۹
فکر می‌کنم گوش‌دادن دانشجویان به مطالب متناقض فقط به سردرگمی و گمراهی آن‌ها منجر خواهد شد.	٤	O٤	علاقه‌ای به تأمل و تفکر جدی درباره سرنوشت و ماهیت جهان یا انسان ندارم.	۱۰	O۱۰
شعر تقریباً اثری بر من ندارد.	٥	O٥	من کنجکاوی فکری فراوانی دارم.	۱۱	O۱۱
اغلب غذاهای جدید و خارجی را امتحان می‌کنم.	٦	O٦	اغلب از کلنجار رفتن با نظریه‌ها یا مفاهیم انتزاعی لذت می‌برم.	۱۲	O۱۲

سؤالات برون‌گرایی					
صورت سؤال	شماره	نماد	صورت سؤال	شماره	نماد
دوست دارم همیشه افراد زیادی دوروبرم باشند.	۱	E۱	اغلب احساس می‌کنم سرشار از انرژی هستم.	۷	E۷
اغلب خود را کمتر از دیگران حس می‌کنم.	۲	E۲	فردی خوشحال و بشاش و دارای روحیهٔ خوبی هستم.	۸	E۸
خودم را فرد خیلی سرحال و سرزنده‌ای نمی دانم.	۳	E۳	شخص بانشاط و خوش‌بینی نیستم.	۹	E۹
واقعاً از صحبت‌کردن با دیگران لذت می‌برم.	٤	E٤	زندگی و رویدادهای آن برایم سریع می‌گذرند.	۱۰	E۱۰
همیشه برای کار آماده‌ام.	٥	E٥	شخص بسیار فعالی هستم.	۱۱	E۱۱
اغلب ترجیح می‌دهم کارها را به‌تنهایی انجام دهم.	٦	E٦	ترجیح می‌دهم برای خودم کار کنم تا راهبر دیگران باشم.	۱۲	E۱۲

سؤالات وظیفه‌شناسی					
صورت سؤال	شماره	نماد	صورت سؤال	شماره	نماد
وسایل متعلق به خود را تمیز و مرتب نگاه می‌دارم.	۱	C۱	برای رسیدن به اهدافم شدیداً تلاش می‌کنم.	۷	C۷
به‌خوبی می‌توانم کارهایم را طوری تنظیم کنم که درست سر زمان تعیین‌شده انجام شوند.	۲	C۲	وقتی قول یا تعهدی می‌دهم، همواره می‌توان برای عمل به آن روی من حساب کرد.	۸	C۸
فرد خیلی مرتب و منظمی نیستم.	۳	C۳	گاهی آن‌طور که بایدوشاید قابل‌اعتماد و اتکا نیستم.	۹	C۹
سعی می‌کنم همه کارهایم را با احساس مسئولیت انجام دهم.	٤	C٤	فرد مولدی هستم که همیشه کارهایم را به اتمام می رسانم.	۱۰	C۱۰
دارای هدف روشنی هستم و برای رسیدن به آن طبق برنامه کار می‌کنم.	٥	C٥	فکر نمی‌کنم که هیچ‌وقت بتوانم فردی منطقی بشوم.	۱۱	C۱۱
قبل از شروع هر کاری وقت زیادی را تلف می‌کنم.	٦	C٦	تلاش می‌کنم هر کاری را به نحو ماهرانه‌ای انجام دهم.	۱۲	C۱۲

منحنی‌های ویژه عملیاتی مربوط به هر سؤال که در قسمت یافته‌ها به آن اشاره شده است در ادامه گزارش شده است.





کدهای دستوری استفاده شده در این پژوهش که در نرم‌افزار Rsudio و تحت پکیج "ltm" اجرا شده است، برای استفاده علاقه‌مندان و پژوهشگران در اینجا ارائه شده است.

```
library("ltm")
library("MASS")
library("msm")
library("polycor")
setwd("D:/ICC")
DATA<- read.table("data.dat", header = T)
Fit١ <- grm(DATA,, constrained = TRUE)
fit2 <- grm(DATA,, constrained = FALSE)
anova(fit1, fit2)
summary(fit2)
plot(fit2, legend = TRUE, cx = "topright", lwd =2, cex = 0.6))
plot(fit2, type = "IIC",items = (1: 12), legend = TRUE, cx = "topright", lwd =2, cex =
0.6)
plot(fit2, type = "OCCu",items = (1: 12), legend = TRUE, cx = "bottomright", lwd
=2, cex = 0.6)
information(fit2, c(-4, 4), items = c(1, 12))
```