University of Sistan and Baluchestan

Journal of Educational Psychology Studies

Print ISSN: 2228 - 6683

The effect of missing data in bias and precision of academic alienation questions

Zahra Jahanbakhsh[1] | **Ali Delavar**[2] | **Noorali Farrokhi**[3] | **Jalil Younesi**[4]

1. Ph.D. Student of Assessment and Measurement, Department of Assessment and Measurement, Faculty of Psychology and Educational Sciences, Allameh Tabataba'I University, Tehran, Iran.. **E-mail:** zjahanbakhsh60@gmail.com
2. **Corresponding Author,** Professor, Department of Assessment and Measurement, Faculty of Psychology and Educational Sciences, Allameh Tabataba'I University, Tehran, Iran. **E-mail:** delavarali@yahoo.com
3. Associate Professor, Department of Assessment and Measurement, Faculty of Psychology and Educational Sciences, Allameh Tabataba'I University, Tehran, Iran. **E-mail:** farrokhinoorali@gmail.com
4. Associate Professor, Department of Assessment and Measurement, Faculty of Psychology and Educational Sciences, Allameh Tabataba'I University, Tehran, Iran. **E-mail:** jalilyounesi@gmail.com

Article Info

Article Type:
Research Article

Received Date:
09 April 2022
Received in Revised From:
20 June 2022
Accepted Date:
23 August 2022
Published Online:
22 September 2022

Keywords:
Missing Data, Academic Alienation, Likelihood Ratio Test, Type I Error Rate, Statistical Power

Abstract

One of the important type of errors is differential item functioning (DIF). The aim of this study, using the Monte Carlo simulation method, was to determine whether the effect of missing data on the bias and precision of the Likelihood Ratio test for DIF detection differ across different missing data methods? The factorial design was used to examine the effect of six MDM methods on the effectiveness of IRT-LR test for DIF detection in the GRM, in terms of Type I error and statistical power. The statistical population in this study was all university students with undergraduate degrees in educational sciences and psychology in Tehran. A sample of 1100 students from this community were among the students studying at city of Tehran such as Allameh Tabataba'i, Tehran, Shahid Beheshti in 2019. The academic alienation questionnaire administered to them. G^2 index and the effect size η^2 calculated for all first-order interactions. Multiple imputation and single regression substitution methods had the lowest contribution in the control of the type I error and the methods based on maximum likelihood and the person mean substitution showed the highest contribution in terms of error control, respectively. The statistical power of the test was similar in all methods in this study. The results of this study indicated that the missing data and observations and the length of the test had an important role in the effectiveness (bias and precision) of the test.

Cite this article: Jahanbakhsh, Z., Delavar, A., Farrokhi, N., & Younesi, J. (2022). The effect of missing data in bias and precision of academic alienation questions. *Journal of Educational Psychology Studies*, 19(46), 14-28.
DOI: 10.22111/JEPS.2022.6757

Publisher: University of Sistan and Baluchestan

مجله مطالعات روانشناسی تربیتی
شاپا چاپی: ۶۶۸۳-۲۲۲۸

تأثیر داده‌های مفقود در سوگیری و دقت سؤالات پرسشنامه‌ی بیگانگی تحصیلی

زهرا جهانبخش[۱] | علی دلاور[۲] | نورعلی فرخی[۳] | جلیل یونسی[٤]

۱. دانشجوی دکتری سنجش و اندازه‌گیری، گروه سنجش و اندازه‌گیری، دانشکده روانشناسی و علوم تربیتی، دانشگاه علامه طباطبائی، تهران، ایران. **رایانامه:** zjahanbakhsh60@gmail.com

۲. **نویسنده مسئول،** استاد ممتاز، گروه سنجش و اندازه‌گیری، دانشکده روانشناسی و علوم تربیتی، دانشگاه علامه طباطبائی، تهران، ایران. **رایانامه:** delavarali@yahoo.com

۳. دانشیار، گروه سنجش و اندازه‌گیری، دانشکده روانشناسی و علوم تربیتی، دانشگاه علامه طباطبائی، تهران، ایران. **رایانامه:** farrokhinoorali@gmail.com

٤. دانشیار، گروه سنجش و اندازه‌گیری، دانشکده روانشناسی و علوم تربیتی، دانشگاه علامه طباطبائی، تهران، ایران. **رایانامه:** jalilyounesi@gmail.com

اطلاعات مقاله

نوع مقاله:
مقاله پژوهشی

تاریخ دریافت:
۱۴۰۱/۰۱/۲۰

تاریخ ویرایش:
۱۴۰۱/۰۳/۳۰

تاریخ پذیرش:
۱۴۰۱/۰۶/۰۱

تاریخ انتشار:
۱۴۰۱/۰۶/۳۱

واژگان کلیدی:
داده مفقود، بیگانگی تحصیلی، آزمون نسبت درست‌نمایی، خطای نوع اوّل، توان آزمون

چکیده

یکی از انواع مهم خطاهای منظم کنش افتراقی است. هدف این پژوهش تعیین این بود که آیا تأثیر داده‌های مفقود بر دقت آزمون نسبت درست‌نمایی برای شناسایی کنش افتراقی در سؤالات چندارزشی در روش‌های مختلف برخورد با داده‌های مفقود متفاوت است؟ تأثیر شش روش برخورد با داده‌های مفقود بر اثربخشی این آزمون برای تشخیص این تفاوت در مدل پاسخ مدرج بر حسب خطای نوع اول و توان آماری بررسی شد. جامعه آماری این پژوهش کلیه دانشجویان دانشگاه‌های دارای رشته‌های علوم تربیتی و روانشناسی مقطع کارشناسی در شهر تهران بودند. ۱۱۰۰ نفر نمونه از بین دانشجویان در حال تحصیل در دانشگاه‌هایی همچون علامه طباطبایی، تهران، شهید بهشتی در سال ۱۳۹۸ بودند که پرسشنامه بیگانگی تحصیلی برای آن‌ها اجرا شد. تحلیل‌ها با استفاده از روش تحلیل رگرسیون چندگانه، مبتنی بر شاخص G^2 و η^2 برای تمامی اثرهای متقابل مرتبه اول محاسبه شد. روش انتساب چندگانه کمترین سهم را در کنترل خطای نوع اول و روش مبتنی بر بیشینه درست‌نمایی بیشترین سهم را نشان داد. توان آزمون در تمامی روش‌های مورد بررسی مشابه بود. نتایج حاکی از این است که نسبت داده‌ها و مشاهدات مفقود و طول آزمون نقش مهمی در تعیین اثربخشی آزمون و به عبارتی سوگیری و دقت سؤالات آزمون داشتند.

استناد به این مقاله: جهانبخش، زهرا؛ دلاور، علی؛ فرخی، نورعلی و یونسی، جلیل. (۱۴۰۱). تأثیر داده‌های مفقود در سوگیری و دقت سؤالات پرسشنامه‌ی بیگانگی تحصیلی. مجله *مطالعات روانشناسی تربیتی*، ۱۹(۴۶)، ۲۸-۱۴.

DOI: 10.22111/JEPS.2022.6757

ناشر: دانشگاه سیستان و بلوچستان

مقدمه

کانون توجه اصلاحات آموزشی ایجاد روش‌هایی است که هدف آن حذف خطاهای منظم است. از آنجایی‌که دقت اندازه گیری لازمه‌ی تفسیری روا است (کرونباخ و گلسر[۱]، ۱۹۶۵) تمرکز پژوهش‌های زیادی بر افزایش روایی[۲] تفسیر از نمرات آزمون‌ها است (رابیچ و روپ[۳]، ۲۰۰۹). به بیان گیرالدو[۴] (۲۰۲۰) روایی مهم‌ترین ملاحظه در تولید و به کارگیری آزمون‌ها است. علی‌رغم پیشرفت مستمر در رواسازی آزمون‌ها، وجود خطاهای منظم اندازه‌گیری تهدیدی برای روایی استنباط از نمرات آزمون است (زومبو[۵]، ۱۹۹۹). یکی از انواع مهم این خطاها زمانی ایجاد می‌شوند که با وجود ثابت نگه داشتن سطح توانایی یا خصیصه، افراد از گروه‌های مختلف (مانند جنسیت) احتمال یکسانی برای ارائه پاسخ صحیح ندارند (بالسیس، گلیسن، وودز و اولتمنز[۶]، ۲۰۰۷؛ امبرتسون و رایس[۷]، ۲۰۰۰). شناسایی کنش افتراقی سؤال[۸] روشی برای شناسایی این نوع خطا است که بُعد مهمی در رواسازی تفسیر نمرات آزمون محسوب می‌شود (گومز–بنیتو، سیرسی، گارسیا و بائینا[۹]، ۲۰۱۷؛ آنکنمن، ویت و دانبار[۱۰]، ۱۹۹۹).

عملکرد افتراقی سؤال زمانی رخ می‌دهد که گروه‌های مختلف آزمون‌شوندگان با توانایی کلی مشابه یا وضعیت مشابه در ملاکی مناسب بطور متوسط پاسخ‌های متفاوت منظمی به یک سؤال خاص ارائه کنند (استانداردهای آزمون تحصیلی و روان‌شناختی[۱۱]، ۲۰۱۴). به این گروه‌ها گروه‌های مرجع (گروه اکثریت)[۱۲] و کانونی (گروه اقلیت یا محروم)[۱۳] می‌گویند (سوامیناتان و راجرز[۱۴]، ۱۹۹۰). سوگیری سؤالات نشان‌دهنده این است که ساختار درونی سؤالات آزمون (ویژگی‌های سؤال، مانند قدرت تمیز و مکان) برای گروه‌های مختلف افراد که توانایی یکسانی دارند مشابه نیست (وودز[۱۵]، ۲۰۰۸؛ رایس، کراینر، دامسگارد و نیلسن[۱۶]، ۲۰۱۸). برای مثال ژو و آریادوست[۱۷] (۲۰۲۰) در پژوهشی به بررسی تأثیر زبان مادری شرکت‌کنندگان در آزمون به این نتیجه رسیدند که سؤالات آزمون تحت تأثیر این متغیر دچار کنش افتراقی هستند. در مطالعه‌ای دیگر، شهمیرزادی، سیری، مرعشی و گرامی پور (۱۳۹۹) نشان دادند تحت مدل جی دینا در مورد

1. Cronbach, & Gleser
2. validity
3. Robitzsch, & Rupp
4. Giraldo
5. Zumbo
6. Balsis, Gleason, Woods, & Oltmanns
7. Embretson, & Reise
8. Differential Item Functioning
9. Gomez-benito, Sireci, García, & Baena
10. Ankenmann, Witt, & Dunbar
11. the Standards for Educational and Psychological Testing
12. reference group
13.focal group
14. Swaminathan, & Rogers
15. woods
16. Rayce, Kreiner, Damsgaard, & Nielsen
17. Zhu, & Aryadoust

آزمون ورودی مقطع دکترای زبان عمومی کنش افتراقی سؤال در گروه دختران و پسران مشاهده شد. با این حال در پژوهش کرمی و خودی (۱۳۹۹) نتایج حاکی از این بود که در بخش دستور زبان و گرامر سؤالات آزمون از روایی مناسبی برخوردار است و کنش افتراقی در سؤالات مشاهده نشد.

تشخیص وجود عملکرد افتراقی در سؤالات چندارزشی در سال‌های اخیر توجه زیادی را به خود جلب کرده است (سو و وانگ[1]، ۲۰۰۵). با وجود این، بیشتر مطالعات در نظریه سؤال پاسخ بر مبنای مدل‌های دو ارزشی بودند (دی آیلا و ساوا–بولستا[2]، ۱۹۹۹) و در حال حاضر نیز شکل دوارزشی آزمون‌های پیشرفت تحصیلی غالباً مورد استفاده قرار می‌گیرد. آزمون کردن با تأکید بر نمره‌گذاری دوارزشی سؤالات می‌تواند آسیب‌رسان باشد؛ بنابراین گزینه‌ای بهتر استفاده از مدل های سنجشی است که اطلاعات را از طبقات مختلف پاسخ اندازه می‌گیرد و این حالتی است که در مدل‌های چندارزشی با آن‌ها مواجه هستیم.

کنش افتراقی سؤال در نظریه سؤال پاسخ به خوبی تعریف شده است، زیرا در این چارچوب پارامترهای سؤال بر اساس عضویت گروهی نامتغیر است (ریو[3]، ۲۰۰۲). سلوی و آلیچی[4] (۲۰۱۸) و پائک و ویلسون[5] (۲۰۱۱) روش‌های شناسایی کنش افتراقی را در چارچوب IRT و نظریه کلاسیک مقایسه کردند و دریافتند که روش‌های IRT توان آماری بیشتر و عملکرد بهتری داشتند. همچنین جین و چن[6] (۲۰۲۰) بیان کردند روش‌های کلاسیک شناسایی DIF همچون رگرسیون لجستیک در سؤالات از نوع لیکرت کارآمد نیستند. در سال‌های اخیر روش‌هایی برای بررسی DIF در زمینه سؤالات چندارزشی گسترش یافته‌اند. از آنجایی که روش آزمون نسبت درست‌نمایی برآورد نااریبی از پارامترها را با داده های دارای مکانیزم داده مفقود کاملاً تصادفی تولید می‌کند (اندرس[7]، ۲۰۱۰)، بنابراین مطالعه حاضر از این روش برای آزمون تشخیص کنش افتراقی سؤال بهره برد. داده‌های مفقود در اغلب پژوهش‌ها به چشم می‌خورند و تهدیدی جدی برای بی‌طرفی در کاربرد آزمون و روایی تفسیر نمرات آزمون هستند. با توجه به میزان و نوع داده‌های مفقود، ممکن است وجود آن به اریبی و خطای برآورد منجر گردد (دمیر[8]، ۲۰۱۳). در شرایطی که میزان داده‌های مفقود زیاد باشد، این موضوع علاوه بر روایی و پایایی نتایج بر تعمیم یافته‌های پژوهش و استنباط‌های آماری آن تأثیر منفی دارد (مک نایت، سیدانی و فیگدرو[9]، ۲۰۰۷).

1. Su, & Wang
2. De Ayala, & Sava-Bolesta
3. Reeve
4. Selvi, & Alici
5. Paek, & Wilson
6. Jin, & Chen
7. Enders
8. Demir
9. McKnight, Sidani, & Figuedero

گیلرا، گومز- بنیتو، هیدالگو و سانچز-مکا[1] (۲۰۱۳) در یک فراتحلیل با بررسی ۴۹ مطالعه در مورد خطای نوع اول و توان آماری آزمون تشخیص DIF بیان کردند که پیامدهای وجود داده‌های مفقود و برخورد با آن‌ها یکی از متغیرهایی است که مطالعه در مورد آن در زمینه آزمون‌های شناسایی DIF مغفول مانده است. شناسایی کنش افتراقی سؤال (DIF) در سؤالات تشریحی با نمره‌گذاری چندارزشی با وجود مشکل داده‌های مفقود ممکن است پیچیده شود، زیرا متغیر مورد اندازه‌گیری و پاسخ حذف شده مستقل نیستند. یکی از این پرسشنامه‌ها که از این طیف سؤالات دارای نمره‌گذاری چند ارزشی استفاده می‌کند، پرسشنامه بیگانگی تحصیلی است. نتایج پژوهش‌ها در ایران نشان‌دهنده بیگانگی تحصیلی دانشجویان از ساختار، فرایند و فرهنگ دانشجویی است (اشرف، شیخ اسلامی، ۱۳۹۶). مورینج، هاجر و هاشر[2] (۲۰۲۰) در مطالعه‌ای طولی افراد را طی سه سال تحصیلی مورد بررسی قرار دادند و دریافتند که افراد در سال سوم تحصیلی بیگانگی تحصیلی بیشتری را نشان دادند. روش‌های آماری که برای ارزیابی DIF به کار می‌روند ممکن است نسبت به داده‌های مفقود مقاوم نباشند و همگرایی رخ ندهد (دراسگو، لوین، ساین، ویلیامز و مید[3]، ۱۹۹۵)؛ بنابراین داده‌ها باید با به‌کارگیری روش‌های برخورد با داده‌های مفقود[4] اصلاح شوند و با آن‌ها به شیوه مناسب برخورد شود. در این مطالعه روش‌های برخورد با داده‌های مفقود شامل روش‌های حذف گام‌به‌گام[5]، مبتنی بر بیشینه درست‌نمایی[6]، انتساب چندگانه[7]، جایگذاری رگرسیون تکی[8]، جایگذاری میانگین نسبی[9] و جایگذاری میانگین شخص[10] بود.

کارکرد آزمون نسبت درست‌نمایی در بستر IRT بر مبنای توان آماری و خطای نوع اول ارزیابی می‌گردد (سدیوی، ژانگ و تراکسل[11]، ۲۰۰۶). در پژوهش‌های پیشین (سلوی و آلیچی، ۲۰۱۸؛ فینچ[12]، ۲۰۱۱؛ بانکز[13]، ۲۰۱۵؛ بانکز و واکر، ۲۰۰۶؛ رابیچ و روپ[14]، ۲۰۰۹) نشان داده شد که وجود داده‌های مفقود و روش‌های برخورد با آن‌ها می‌تواند در توانایی روش‌های رایج شناسایی DIF مؤثر باشد. در برخی موارد خطای نوع اول روش‌های شناسایی DIF متورم شده بود (بانکز و واکر[15]، ۲۰۰۶؛ رابیچ و روپ، ۲۰۰۶) و در برخی دیگر توان آزمون در شناسایی DIF در حضور داده‌های مفقود کم بود (سدیوی، ژانگ و تراکسل، ۲۰۰۶؛ سیجسما و وندر آرک[16]، ۲۰۰۳).

1. Guilera, Gómez-Benito, Hidalgo, & Sánchez-Meca
2. Morinaj, Hadjar, & Hascher
3. Drasgow, Levine, Tsien, Williams, & Mead
4.missing data method
5. Listwise deletion
6. full-information maximum likelihood
7. multiple imputation
8. single regression substitution
9. relative mean substitution
10. person mean substitution
11. Sedivy, Zhang, & Traxel
12. Finch
13. Banks
14.Robitzsch, & Rupp
15. Banks, & Walker
16. Sijtsma,, & van der Ark

پژوهش‌ها نشان دادند ممکن است عوامل مداخله‌گر دیگری اثرگذاری روش‌های برخورد با داده‌های مفقود بر نرخ خطای نوع اول و توان آماری آزمون تشخیص کنش افتراقی سؤال را دستخوش تغییر سازند (هیدالگو، لوپز-مارتینز، گوم- بنیتو و گیلرا[1]، ۲۰۱۶؛ فینچ، ۲۰۱۱؛ رابیچ و روپ، ۲۰۰۹؛ گرت[2]، ۲۰۰۹؛ وانگ و یه[3]، ۲۰۰۳). هیدالگو و همکاران (۲۰۱۶) در پژوهشی به مقایسه روش رگرسیون لجستیک و آزمون نسبت درست‌نمایی مبتنی بر IRT در مورد سؤالات چند ارزشی پرداختند. آن‌ها دریافتند که خطای نوع اول تحت تأثیر طول آزمون، شدت DIF، اندازه نمونه و تعداد طبقات پاسخ بود.

این پژوهش منطبق بر مدل پاسخ مدرج (GRM؛ سامیجیما[4]، ۱۹۶۹، ۲۰۱۰) به دنبال پاسخ به این سؤال‌های پژوهشی بود: ۱) آیا تأثیر داده‌های مفقود در خطای نوع اول آزمون نسبت درست‌نمایی (IRT-LR) برای شناسایی DIF در سؤالات چندارزشی در روش‌های مختلف برخورد با داده‌های مفقود متفاوت است؟ و ۲) آیا تأثیر داده‌های مفقود در توان آماری آزمون نسبت درست‌نمایی (IRT-LR) برای شناسایی DIF در سؤالات چندارزشی در روش‌های مختلف برخورد با داده‌های مفقود متفاوت است؟

روش

این مطالعه از روش شبیه‌سازی مونت کارلو با استفاده از پارامترهای داده‌های پرسشنامه بیگانگی تحصیلی که توسط پژوهشگر گردآوری شد، داده‌های سؤال پاسخ را به نحوی تولید کرد تا با GRM مطابقت داشته باشد. از طرح عاملی برای ارزیابی تأثیر شش روش برخورد با داده مفقود بر اثربخشی IRT-LR در مدل پاسخ مدرج (GRM) بر حسب خطای نوع اول و توان آماری استفاده شد (جدول ۱).

جدول ۱. شرایط مطالعه شبیه‌سازی

عوامل شبیه‌سازی شده
روش‌های برخورد با داده مفقود
اندازه نمونه
طول آزمون
میزان افراد مفقود
میزان سؤالات مفقود
شدت DIF
توزیع توانایی گروه کانونی و گروه مرجع

برای تولید داده‌ها جهت شبیه‌سازی، از مجموعه داده‌های حقیقی استفاده شد تا عوامل و روش‌های مورد نظر در مورد موقعیت‌های واقعی به کار گرفته شوند. بدین ترتیب نمونه‌گیری در دو بخش انجام شد:

1. Hidalgo, López-Martínez, Gómez-Benito, & Guilera
2. Garret
3. Wang, & Yeh
4. Samejima

الف) در بخش اول که مربوط به برآورد پارامترها بود، جامعه آماری کلیه دانشجویان دانشگاه‌های دارای رشته‌های علوم انسانی ایران بودند. جامعه آماری در این پژوهش کلیه دانشجویان دانشگاه‌های دارای رشته‌های علوم تربیتی و روانشناسی مقطع کارشناسی در شهر تهران بودند. نمونه‌ی مورد پژوهش از این جامعه از بین دانشجویان در حال تحصیل در دانشگاه‌های علامه طباطبایی، تهران، شهید بهشتی، خوارزمی، آزاد اسلامی واحد تهران شمال، آزاد اسلامی تهران جنوب و آزاد اسلامی تهران مرکز انتخاب شدند. بدین منظور جهت اجرای پرسشنامه، به دانشگاه‌های دارای رشته‌های علوم تربیتی و روانشناسی مراجعه شد و پس از اینکه هدف از انجام این مطالعه به دانشجویان توضیح داده شد، از بین دانشجویان علاقه‌مند به همکاری ۱۱۰۰ نفر به روش غیرتصادفی (هدفمند) انتخاب شدند و برای پاسخگویی به سؤالات مقیاس بیگانگی تحصیلی مورد پرسش واقع شدند. ۴۲۷ نفر (۳۹٪) از شرکت‌کنندگان پسر و ۶۷۳ نفر (۶۱٪) از آن‌ها دختر بودند. برای جمع‌آوری داده‌ها از ابزار زیر استفاده شد.

داده‌های این پژوهش با استفاده از پرسشنامه بیگانگی تحصیلی (بورباچ، ۱۹۷۲) به صورت طیف لیکرت با مقیاس ۴ درجه‌ای کاملاً مخالفم (۱)، تاحدی مخالفم (2)، تاحدی موافقم (3)، کاملاً موافقم (۴) به دست آمد. این پرسشنامه دارای سه زیرمقیاس انزوا (۴ سؤال)، بی‌معنایی (۵ سؤال)، ناتوانی (۶ سؤال) بود. به گزارش بورباچ[1] (۱۹۷۲) پایایی همسانی درونی این مقیاس در بین دانشجویان سال دوم ۰/۹۲ بود. به‌علاوه برای زیر مقیاس ناتوانی، بی‌معنایی و انزوا ثبات درونی به ترتیب ۰/۸۴، ۰/۸۶ و ۰/۷۲ بود. نتایج بررسی پایایی همسانی درونی و روایی سازه زیرمقیاس‌ها در مطالعه حاضر در جدول ۲ خلاصه شد:

جدول ۲. نتایج مربوط به همسانی درونی و روایی سازه

زیرمقیاس	آلفای کرونباخ	CFI	GFI	AGFI	RMSEA
انزوا	۰/۷۸	۰/۹۶	۰/۹۲	۰/۸۴	۰/۰۶۱
بی‌معنایی	۰/۸۴	۰/۹۵	۰/۹۰	۰/۸۸	۰/۰۵۷
ناتوانی	۰/۸۴	۰/۹۱	۰/۹۰	۰/۸۶	۰/۰۶۴

پاسخ‌های شرکت‌کنندگان به پرسشنامه بیگانگی تحصیلی از طریق حضور پژوهشگر در دانشکده‌های روانشناسی و علوم تربیتی دانشگاه‌های تهران و توزیع پرسشنامه‌ها به صورت جمعی بین دانشجویان مقطع کارشناسی گردآوری شد. پس از اینکه هدف از انجام پژوهش به شرکت‌کنندگان توضیح داده شد، از افراد علاقه‌مند به همکاری درخواست شد تا پاسخ به پرسشنامه را آغاز کنند. لازم به ذکر است که در ابتدای جلسه بیان شد هیچ یک از سؤالات دارای پاسخ صحیح یا غلط نیستند و افراد صرفاً گزینه‌ای را انتخاب کنند که متناسب با نگرش شخصی آن‌ها نسبت به سؤال مورد نظر است و تا حد ممکن هیچ سؤالی را بدون پاسخ نگذارند.

ب) در بخش دوم که مربوط به شبیه‌سازی داده‌ها بود، روش شبیه‌سازی به عنوان روش گردآوری داده‌ها کمک کرد تا از این جامعه موقعیت‌های متعددی را ایجاد کرد و بتوان به الگوهای متنوعی از بی‌پاسخی دست پیدا کرد. دو حجم

1. Burbach

نمونه کل با توجه به پارامترهای حاصل از داده‌هایی که از بخش اول به دست می‌آید، شبیه‌سازی شد. یک حجم نمونه کل بزرگ (N_1=۱۰۰۰) که امکان برآوردهایی مناسب را داشته باشد و یک حجم نمونه کل حداقلی (N_2 =۵۰۰) برای به‌کارگیری GRM شبیه‌سازی شد. در پژوهش‌های DIF به طور معمول نمونه‌های نامتوازن را به کار می‌گیرند که در آن‌ها نمونه کانونی کوچک‌تر از نمونه مرجع است؛ بنابراین علاوه بر نمونه‌های متوازن[1] با نسبت ۱:۱ و نمونه‌های نامتوازن با نسبت ۳:۲ برای گروه‌های مرجع و کانونی تولید شد. مقادیر پارامترهای سؤال به‌وسیله نرم‌افزار MULTILOG 7.03 مدرج شدند. داده‌های سؤال پاسخ برای مدل پاسخ دو پارامتری مدرج چندارزشی تولید شد.

با استفاده از نرم‌افزار R 4.0.3 داده‌های سؤال پاسخ شبیه‌سازی شده برای تحلیل DIF با توزیع نرمال (۱ و ۰)N برای سه طول آزمون (۴، ۵ و۶ سؤالی) با ۴ طبقه پاسخ از نوع لیکرت برای ۱۰۰۰ نمونه (تکرار[2]) برای هر یک از شرایط شبیه‌سازی تولید شد. شبیه‌سازی داده‌های سؤال پاسخ برای مدل پاسخ مدرج دو پارامتری چندارزشی با استفاده از معادله داد، دی آیلا و کوچ[3] (۱۹۹۵) محاسبه شد.

تأثیر روش‌های برخورد با داده‌های مفقود بر اثربخشی آزمون IRT-LR برای تشخیص DIF برای هر یک از ترکیب‌های عامل‌ها بر مبنای خطای نوع اول و توان آماری سنجیده شد. به‌کارگیری آزمون IRT-LR برای شناسایی DIF مستلزم آزمون فرض صفر از طریق مقایسه مدل تکمیلی[4] در مقابل مدل پایه[5] است تا مشخص شود آیا پارامترهای افزوده در مدل تکمیلی به طور معناداری متفاوت هستند (تیسن و همکاران، ۱۹۹۳). نیرومندی آزمون نسبت درست‌نمایی IRT برای کنترل خطای نوع اول در سطح معناداری ۰/۰۵ بررسی شد. از آنجایی که رویکرد خط پایه به IRT-LR برای تشخیص DIF مستلزم مجموعه‌ای از مقایسه‌ها است، تصحیح بونفرونی در مورد سطح معناداری اعمال شد تا از تورم آلفا جلوگیری شود. تصمیم برای رد فرض صفر بر اساس مقایسه آماره G^2 با آماره χ^2 و مقادیر بحرانی بونفرونی اتخاذ شد. برای تحلیل کل تغییرات نرخ خطای آزمون نسبت درست‌نمایی که توسط تعامل آن با عامل‌ها تبیین می‌شود، اندازه اثر η^2 برای تمامی اثرهای متقابل مرتبه اول با روش محاسبه شد تا تعیین شود کدام یک از تعامل‌ها سهم معناداری دارند ($\eta^2 \geq$ ۰/۰۵). سپس توان آزمون نسبت درست‌نمایی برای هر یک از آزمون‌ها برآورد شد. تحلیل توان آزمون فقط برای حالاتی که کنترل خطای نوع اول آن‌ها بر اساس ملاک برادلی[6] (۱۹۷۸) کافی است انجام شد.

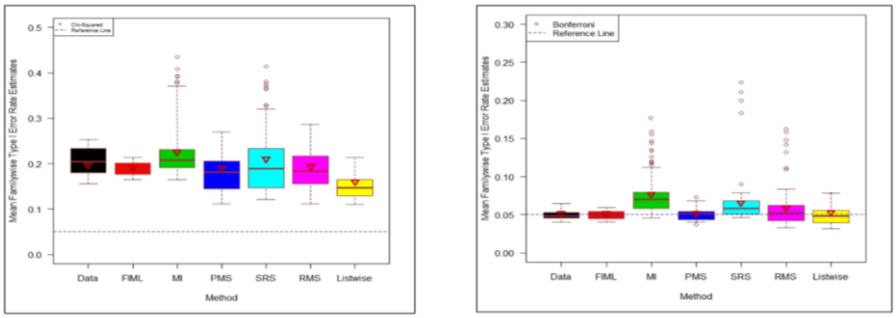
یافته‌ها

این مطالعه تأثیر شش روش برخورد با داده‌های مفقود را بر عملکرد آزمون IRT-LR برای تشخیص کنش افتراقی سؤال (DIF) در سؤالات دارای نمره‌گذاری چندارزشی بر مبنای خطای نوع اول و توان آماری مقایسه کرد. برای پاسخ به

1. balanced
2. replication
3. Dad, De Ayla, Coch
4. augmented model
5. baseline model
6. Bradley

سؤال پژوهش، ابتدا توزیع خطای نوع اول در هر یک از روش‌های برخورد با داده‌های مفقود در تمام شرایط شبیه‌سازی برای هر دو مقدار بحرانی خی دو و تصحیح بونفرونی بر مبنای میانگین خانواده نرخ‌های خطا[1] در سطح معناداری ۰/۰۵ بررسی شد.

همان‌گونه که در نمودار ۱ دیده می‌شود، در همه روش‌ها توزیع برآوردهای میانگین خانواده خطا دارای مقدار میانگین M =۰/۱۶ تا M =۰/۲۳ است ($\chi^2_{(۴)}$ =۹/۴۹). استفاده از χ^2 برای آزمون باعث تورم میانگین خطا شد. با استفاده از تصحیح بونفرونی، میانگین خطا تقریباً برابر با آلفای اسمی (۰/۰۵) یا اندکی بالاتر بود. همچنین روش انتساب چندگانه و جایگذاری رگرسیون تکی دارای میانگین بالاتری از سایر روش‌ها بودند (به ترتیب M =۰/۰۸ و M =۰/۰۷). مقادیر پرت بالایی در سه روش انتساب چندگانه، جایگذاری رگرسیون تکی و جایگذاری میانگین نسبی بیانگر تأثیر عامل یا عوامل مورد بررسی است. مقادیر اندازه اثر تنها برای تصحیح بونفرونی محاسبه شد تا اثرهای متقابل و اصلی برای هر یک از روش‌ها و عوامل تعیین شود.



نمودار ۱. توزیع کلی خانواده نرخ‌های خطا بر اساس روش جایگذاری داده‌های مفقود به ترتیب با خی دو بحرانی و تصحیح بونفرونی (α =۰/۰۵). خط افقی نمایانگر سطح α است.

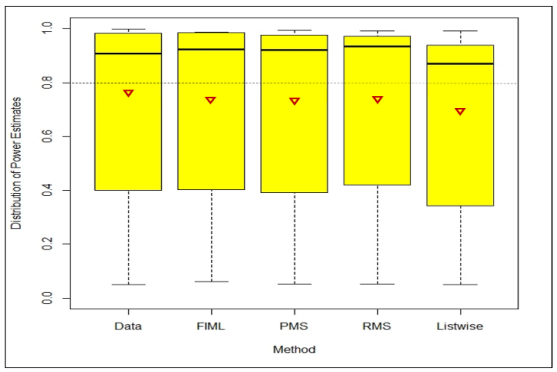
اندازه‌های اثر برای همه روش‌ها محاسبه شد تا عوامل مؤثر در تغییرات میانگین خانواده خطا شناسایی شود. در روش جایگذاری رگرسیون تکی تحلیل اندازه اثر برای اثرهای متقابل، اثر بزرگی برای سه اثر نشان داد: ۱) اثر متقابل نسبت مشاهدات مفقود و نسبت سؤالات مفقود (η^2 =۰/۱۲)؛ ۲) اثر متقابل تعداد سؤالات و نسبت مشاهدات مفقود (η^2 =۰/۰۸)؛ ۳) اثر متقابل تعداد سؤالات و نسبت سؤالات مفقود (η^2 =۰/۱۱). نتایج برای روش انتساب چندگانه نشان داد اثر متقابل نسبت مشاهدات مفقود و نسبت سؤالات مفقود بزرگ بود (η^2 =۰/۱۰). توزیع میانگین خانواده خطا برای روش جایگذاری میانگین نسبی نشان داد که اثرهای متقابل برای حالات زیر معنادار است: ۱) اثر متقابل تعداد سؤالات و نسبت سؤالات مفقود (η^2 =۰/۱۴)؛ ۲) اثر متقابل تعداد سؤالات و نسبت مشاهدات مفقود (η^2 =۰/۱۰)؛ ۳) اثر متقابل نسبت مشاهدات مفقود و نسبت سؤالات مفقود (η^2 =۰/۰۸). توزیع مربوط به روش حذف گام‌به‌گام نشان داد در این روش تنها اثر متقابل اندازه نمونه و نسبت مشاهدات مفقود معنادار شد (η^2 =۰/۰۷).

1. Familywise error rate

بر مبنای ملاک برادلی از آنجایی که خطای نوع اول در استفاده از χ^2 برای آزمون فرض صفر متورم شد، ملاک برادلی برای نیرومندی برای سطح معناداری ۰/۰۵ برآورده نشد؛ بنابراین تنها شرایطی در تصحیح بونفرونی که حائز ملاک برادلی بودند بر اساس روش‌ها و عوامل مورد پژوهش گزارش شد.

روش انتساب چندگانه و جایگذاری رگرسیون تکی در مقایسه با سایر روش‌ها کمترین سهم از کنترل خطای نوع اول را در بین همه عوامل دارا بودند و از تحلیل‌های بعدی در توان آماری خارج شدند. از سوی دیگر روش‌های داده‌های کامل، روش مبتنی بر بیشینه درست‌نمایی و جایگذاری میانگین شخص به ترتیب بیشترین سهم را از نظر دارا بودن ملاک برادلی نشان دادند.

برای پاسخ به سؤال دوم پژوهش، مقایسه‌ی توان آزمون برای روش‌هایی انجام شد که دارای کنترل کافی در خطای نوع اول بودند. نتایج نشان داد تفاوت در میانگین توان آماری روش‌ها نسبتاً کوچک شد و کمترین مقدار آن ۰/۷۰ در روش حذف گام‌به‌گام بود. این نمودار توزیع برآورد توان آماری روش‌هایی که دارای کنترل کافی بر خطای نوع I هستند را نشان داد. با توجه به نمودار ۲ توزیع برآورد توان آماری روش‌ها در سطح ۰/۰۵ به لحاظ پراکندگی توزیع‌ها دارای کشیدگی منفی بود و در همه روش‌ها میانه بالاتر از ۰/۸ بود. همچنین طول نیمه پایینی توزیع در مقایسه با نیمه بالایی آن بیشتر بود.



نمودار۲. توزیع کلی برآورد توان بر اساس روش‌های جایگذاری داده‌های مفقود در شرایط شبیه‌سازی شده

تحلیل اندازه اثر نشان داد که برای همه روش‌ها اثر متقابل تعداد سؤالات و اندازه نمونه معنادار بود (η^2 =۰/۰۶ تا η^2 =۰/۱۴). در داده‌های کامل، روش‌های مبتنی بر بیشینه درست‌نمایی و جایگذاری میانگین شخص، اثر متقابل تعداد سؤالات و اندازه نمونه اثر معناداری بر میانگین توان آماری آزمون داشت (η^2 =۰/۱۴؛ η^2 =۰/۱۲؛ η^2 =۰/۱۱). در روش جایگذاری میانگین نسبی اثر متقابل اندازه نمونه و تعداد سؤالات اثر متوسطی بر برآورد میانگین توان آماری داشت (η^2 =۰/۰۹). در همه روش‌ها توان آماری در نمونه‌های با اندازه بزرگتر به طور پیوسته بالاتر بود.

بحث و نتیجه‌گیری

هدف کلی این پژوهش مقایسه تأثیر داده‌های مفقود بر خطای نوع اول و توان آماری آزمون IRT-LR برای شناسایی DIF در سؤالات چندارزشی بود. مدل پاسخ مدرج (GRM) در چارچوب نظریه سؤال پاسخ مدل آماری با کفایتی برای نوع داده‌های مورد بررسی در این مطالعه بود. علاوه بر تحلیل داده‌های کامل، شش روش برخورد با داده‌های مفقود حاصل از سؤالات مقیاس سنجش نگرش بیگانگی تحصیلی دانشجویان بر مبنای شبیه‌سازی سطوح مختلف عوامل مورد مطالعه انجام شد. نتایج حاصل از این مطالعه در پاسخ به دو پرسش پژوهشی بود: ۱) آیا تأثیر داده‌های مفقود در خطای نوع اول آزمون نسبت درست‌نمایی برای شناسایی DIF در سؤالات چندارزشی در روش‌های مختلف برخورد با داده‌های مفقود متفاوت است؟ و ۲) آیا تأثیر داده‌های مفقود در توان آماری آزمون نسبت درست‌نمایی برای شناسایی DIF در سؤالات چندارزشی در روش‌های مختلف برخورد با داده‌های مفقود متفاوت است؟

با بررسی توزیع میانگین خطای نوع اول، نتایج این پژوهش در مورد خطای نوع I برای داده‌های کامل، روش مبتنی بر بیشینه درست‌نمایی، روش جایگذاری میانگین شخص تقریباً یکسان بود و تغییر کمی در توزیع خطای نوع اول در همه عوامل مورد بررسی مشاهده شد. تحلیل اندازه اثر برای این روش‌ها نشان داد که در داده‌های کامل هیچ یک از عوامل اثر معناداری بر خطای نوع اول ندارند؛ این نتایج با یافته‌های مطالعه فینچ (۲۰۱۱) سازگار است. در مطالعه فینچ (۲۰۱۱) و گرت (۲۰۰۹)، نتایجی که در دو مطالعه فوق در مورد ثبات خطای اندازه‌گیری در شرایط طرح پژوهش آن‌ها حاصل شد، با پژوهش حاضر سازگار است. این دو مطالعه در بررسی خطای نوع اول در داده‌های کامل، با استفاده از آزمون‌های M-H، SIBTEST و OLR نتیجه گرفتند نرخ خطای نوع اول تقریباً برابر با سطح آلفای اسمی (۰/۰۵) بود و اندازه‌های اثر نیز برای هیچ یک از عوامل مورد بررسی در مطالعه آن‌ها معنادار نبود. مشابه با نتایج پژوهش حاضر، در روش حذف گام‌به‌گام اثر متقابل اندازه نمونه و نسبت مشاهدات مفقود بر نرخ خطای نوع اول اثر داشت.

چنانکه در نتایج گزارش شد، روش انتساب چندگانه دارای بزرگ‌ترین میانگین خطا بود. از سویی دیگر اثر متقابل نسبت مشاهدات مفقود و نسبت سؤالات مفقود با افزایش تعداد سؤالات مفقود افزایش یافت؛ این نتایج با نتایج پژوهش گرت (۲۰۰۹) که حاکی از کاهش نرخ خطا بود ناسازگار است. فینچ (۲۰۱۱) گزارش کرد که با استفاده از روش انتساب چندگانه میزان خطا نزدیک به سطح آلفای اسمی بود که این نتیجه با یافته‌های پژوهش حاضر ناسازگار است؛ مطالعه فینچ با داده‌های دوارزشی و تعداد سؤالات زیاد انجام شده بود. به‌طورکلی در پژوهش حاضر، سازگار با پژوهش‌های قبلی (گرت، ۲۰۰۹؛ رابیچ و روپ، ۲۰۰۹؛ سلوی و ازدمیر آلیجی، ۲۰۱۷) نتایج مربوط به خطای نوع اول در روش‌های مختلف برخورد با داده‌های مفقود و تحت عوامل مورد پژوهش مشابه بود و با نتایج پژوهش سلوی و ازدمیر آلیجی (۲۰۱۷) که بر استفاده از روش‌های کلاسیک شناسایی DIF تأکید داشت، ناسازگار بود. در این پژوهش که با کاربرد آزمون IRT-LR انجام شد، متورم شدن نرخ خطا و اثر بزرگ متقابل بین نسبت مشاهدات مفقود و نسبت داده‌های مفقود بیانگر این بود که روش‌های انتساب چندگانه و جایگذاری رگرسیون تکی کم اثرترین روش‌ها برای برخورد با داده‌های مفقود در سؤالات از نوع لیکرت در مقیاس‌های با طول کوتاه بودند.

نتایج نشان داد توان آماری آزمون IRT-LR در تمامی روش‌های برخورد با داده‌های مفقود و همه عوامل مورد بررسی در این پژوهش مشابه بود. در همه روش‌های برخورد با داده‌های مفقود، اثر متقابل اندازه نمونه و تعداد سؤالات مقیاس، تنها عاملی بود که در این مطالعه اثر قابل توجهی در میزان برآورد توان آزمون داشت. این نتایج با پژوهش‌های قبلی (فینچ، ۲۰۱۱؛ گرت، ۲۰۰۹) کاملاً سازگار بود. در نهایت فارغ از تغییر در نسبت داده‌های مفقود، با کاهش حجم نمونه، توان آماری آزمون کاهش یافت.

در برخورد با داده‌های مفقود، انتخاب روش مناسب باید متناسب با مسئله مورد پژوهش انجام پذیرد. در مقیاس‌های کوتاه، نسبت مقادیر مفقود در متغیرها عامل تأثیرگذاری در کارکرد روش‌های شناسایی DIF هستند. با توجه به پژوهش حاضر، استفاده از روش مبتنی بر بیشینه درست‌نمایی به عنوان بهترین روش برخورد با داده‌های مفقود پیشنهاد می‌گردد. با وجود اینکه در ادبیات پژوهشی حمایت ویژه‌ای از روش انتساب چندگانه وجود دارد؛ با توجه به بررسی اختصاصی داده های از نوع لیکرت در پژوهش حاضر پیشنهاد می‌گردد در به کارگیری این روش در برخورد با داده‌های مفقود در مقیاس های از نوع لیکرت احتیاط شود. از جهت اینکه این روش مقدار برآورد شده را جایگزین مقدار مفقود می‌نماید و این مقادیر محتمل گرد می‌شوند، می‌توانند سبب ایجاد اریبی در برآوردها شوند. علاوه بر مکانیسم داده‌های مفقود کاملاً تصادفی، مکانیسم‌های دیگری چون مفقود به صورت تصادفی و مفقود به صورت غیرتصادفی نیز وجود دارد، برای پژوهش‌های آتی پیشنهاد می‌گردد این نوع داده‌های مفقود نیز مورد بررسی قرار گیرند. در مقایسه با روش‌های کلاسیک مانند منتل هنزل، رگرسیون لجستیک که اثربخشی آن‌ها برای شناسایی کنش افتراقی سؤال نشان داده شده است، روش آزمون نسبت درست‌نمایی در چارچوب نظریه سؤال پاسخ به دلیل پیچیدگی به کارگیری آن به ندرت مورد استفاده قرار می‌گیرد. نرم‌افزاری که برای اجرای آزمون نسبت درستنمایی به کار گرفته می‌شود (برنامه‌نویسی در بستر R یا SAS) اختصاصی‌تر از پکیج‌های آماری رایج است. با وجود اینکه مطالعات اخیر به مدل‌های نظریه پاسخ در این نرم‌افزارها پرداخته‌اند، به کارگیری آزمون نسبت درست‌نمایی در چارچوب نظریه سؤال پاسخ همچنان به ندرت صورت می‌پذیرد و به همین دلیل تعمیم نتایج را با محدودیت مواجه می‌سازد.

این مقاله برگرفته از رساله دکتری با عنوان: "تأثیر داده‌های مفقود بر خطای نوع اوّل و توان آزمونِ نسبت درست نمایی برای تشخیص کنش افتراقی سئوال (DIF)" است.

منابع

آلن، مری؛ ین، وندی. (۱۹۷۹). *مقدمه‌ای بر نظریه‌های اندازه‌گیری*، ترجمه علی دلاور (۱۳۹۲)، تهران: انتشارات سمت.

اشرف، پریچهر؛ شیخ‌الاسلامی، راضیه. (۱۳۹۶). بیگانگی تحصیلی در دانش آموزان: نقش کنترل روانی والدین و تأکیدات هدفی معلم. *مجله مطالعات روانشناسی تربیتی*. ۱۴(۲۷). ۲۹-۶۲.

تیلور، کاترین. (۲۰۱۳). *روایی و رواسازی*، ترجمه جلیل یونسی (۱۳۹۸). تهران: انتشارات علامه طباطبایی.

شهمیرزادی، نیلوفر؛ سیری، مسعود؛ مرعشی، حمید و گرامی پور، مسعود. (۱۳۹۹). بررسی سوگیری در سؤالات درک مطلب آزمون مقطع دکترای رشته زبان انگلیسی تحت سنجش تشخیصی شناختی. *پژوهش‌های زبان‌شناختی در زبان‌های خارجی*، ۱۰(۱)، ۱۵۲-۱۶۵.

کرمی، حسین؛ خودی، علی. (۱۳۹۹). بررسی کنش افتراقی پرسش‌ها و عملکرد در آزمون: مقایسه رگرسیون لجستیک، مدل رش و منتل- هنزل. *پژوهش‌های زبان‌شناختی در زبان‌های خارجی*، ۱۰(۴)، ۸۴۲-۸۵۳.

References

Ankenmann, R. D., Witt, E. A., & Dunbar, S. B. (1999). An investigation of the power of the likelihood ratio goodness-of-fit statistic in detecting differential item functioning. *Journal of Educational Measurement*, 36(4), 277-300.

Balsis, S., Gleason, M. E., Woods, C. M., & Oltmanns, T. F. (2007). An IRT analysis of DSM-IV personality disorder criteria across younger and older age groups. *Psychology and aging*, 22(1), 171-185.

Banks, K. (2015). An introduction to missing data in the context of differential item functioning. *Practical Assessment, Research, and Evaluation*, 20(1), 12-35.

Banks, K., & Walker, C. (2006). *Performance of SIBTEST when focal group examinees have missing data*. San Francisco: National Council of Measurement in Education.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2), 144-152.

Buhi, E. R., Goodson, P., Neilands, T. B. (2008). Out of sight, not out of mind: Strategies for handling missing data. *American Journal of Health Behavior*, 32(1), 83-92.

Burbach, H. J. (1972). An empirical study of powerlessness among high school students. *The High School Journal*, 55(7), 343-354.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth, Thompson Learning.

Demir, E. (2013). Item and test parameters estimations for multiple-choice tests in the presence of missing data: The case of SBS. *Journal of Educational Sciences Research*, 3(2), 47–68.

Drasgow, F., Levine, M. V., Tsien, S., Williams, B., & Mead, A. D. (1995). Fitting polytomous Item Response Theory models to multiple-choice tests. *Applied Psychological Measurement*, 19(2), 143-165.

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

Enders, C. K. (2013). Dealing with missing data in developmental research. *Child Development Perspectives*, 7(1), 27-31.

Finch, H. (2011). The use of multiple imputation for missing data in uniform DIF analysis: Power and Type I error rates. *Applied Measurement in Education*, 24(4), 281-301.

Garrett, P. L. (2009). *A Monte Carlo study investigating missing data, DIF, and effect*. (Doctoral dissertation). Georgia State University, Atlanta, GA. Retrieved from ProQuest LLC. (UMI 3401601).

Gelin, M. N., & Zumbo, B. D. (2007). Operating characteristics of the DIF MIMIC approach using Jöreskog's covariance matrix with ML and WLS estimation for short scales. *Journal of Modern Applied Statistical Methods*, 6(2), 22.

Giraldo, F. (2020). Validity and Classroom Language Testing: A Practical Approach. *Colombian Applied Linguistics Journal*, 22(2), 194-206.

Guilera, G., Gómez-Benito, J., Hidalgo, M. D., & Sánchez-Meca, J. (2013). Type I error and statistical power of the Mantel-Haenszel procedure for detecting DIF: A meta-analysis. *Psychological Methods*, 18(4), 553.

Gómez-Benito, J., Balluerka, N., González, A., Widaman, K. F., & Padilla, J. L. (2017). Detecting differential item functioning in behavioral indicators across parallel forms. *Psicothema*, 29(1), 91-95.

Gómez-Benito, J., Sireci, S., García, J. L. P., Montesinos, M. D. H., & Baena, I. B. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema*, 30(1), 104-109.

Hidalgo, M. D., López-Martínez, M. D., Gómez-Benito, J., & Guilera, G. (2016). A comparison of discriminant logistic regression and Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning (IRTLRDIF) in polytomous short tests. *Psicothema*, 83-88.

Jin, K., Chen H. (2020). MIMIC approach to assessing DIF with control of extreme response style. *Behavior Research Methods*, 52(1)- 131-147.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figuedero, A. J. (2007). *Missing data: A gentle introduction*. New York, NY: Guilford Press.

Morinaj, J., Hadjar, A., & Hascher, T. (2020). School alienation and academic achievement in Switzerland and Luxembourg: a longitudinal perspective. *Social psychology of education*, 47(3), 1-36.

O'Rouke, T. W. (2003). Methodological techniques for dealing with missing data. *American Journal of Health Studies*, 18(2/3), 165-168.

Paek, I., & Wilson, M. (2011). Formulating the Rasch differential item functioning model under the marginal maximum likelihood estimation context and its comparison with Mantel–Haenszel procedure in short test and small sample conditions. *Educational and Psychological Measurement*, 71(6), 1023-1046.

Rayce, S. B., Kreiner, S., Damsgaard, M. T., Nielsen, T., & Holstein, B. E. (2018). Measurement of alienation among adolescents: construct validity of three scales on powerlessness, meaninglessness and social isolation. *Journal of patient-reported outcomes*, 2(1), 1-12.

Reeve, B. B. (2002). An introduction to modern measurement theory. *National Cancer Institute*, 1-67.

Robitzsch, A., & Rupp, A. (2009). Impact of missing data on the detection of differential item functioning: The case of Mantel-Haenszel and logistic regression analysis. *Educational and Psychological Measurement*, 69(1), 18-34.

Samejima, F. (2010). *The general graded response model*. In M. L. Nering & R. Ostini (Eds.), Handbook of polytomous item response theory (pp. 77-107), New York, NY: Routledge, Taylor & Francis Group.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7(2), 147-177.

Scott, N. W., Fayers, P. M., Aaronson, N. K., Bottomley, A., de Graeff, A., Groenvold, M., & Sprangers, M. A. (2010). Differential item functioning (DIF) analyses of health-related quality of life instruments using logistic regression. *Health and quality of life outcomes*, 8(1), 1-9.

Sedivy, S. K., Zhang, B., & Traxel, N. M. (2006). *Detection of differential item functioning with polytomous items in the presence of missing data*. California: National Council of Measurement in Education.

Selvi, H., & Alici, D. Ö. (2018). Investigating the impact of missing data handling methods on the detection of DIF. *International Journal of Assessment Tools in Education*, 5(1), 1-14.

Thissen, D. (2003). *MULTILOG 7.03 User's guide: Multiple categorical item analysis and test scoring using item response theory*. Mooresville, IN: Scientific Software International.

Wang, W. C., & Yeh, Y. L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, 27(6), 479-498.

Widaman, K. F. (2006). Missing data: What to do with or without them. *Monographs of the Society for Research in Child Development*, 7(13), 42-64.

Woods, C. M. (2008). Likelihood-ratio DIF testing: Effects of non-normality. *Applied Psychological Measurement*, 32(7), 511-526.

Zhu, X., & Aryadoust, V. (2020). An investigation of mother tongue DIF in a high –stakes computerized academic reading test. *Computer Assisted Language Learning*, 40(3), 1-25.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.