Educational Measurement and Evaluation Studies

National Organization for Educational Testing

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

The Effect of the Anchor to Total Test Correlation on Equating Results: A Systematic Review

Vahideh Asadi [1], Ali Moghadamzadeh [2], Keyvan Salehi [3]

1. PhD Student of Educational Measurement and Evaluation, Faculty of Psychology and Education, University of Tehran, Tehran, Iran; (Corresponding Author), Email: vahideh.asadi@ut.ac.ir
2. Associate Professor, Division of Research and Assessment, Faculty of Psychology and Education, University of Tehran, Tehran, Iran. Email: amoghadamzadeh@ut.ac.ir
3. Associate Professor, Division of Research and Assessment, Faculty of Psychology and Education, University of Tehran, Tehran, Iran. Email: keyvansalehi@ut.ac.ir

Article Info

Article Type:
Research Article

Received: 2023.03.27
Received in revised form: 2023.06.26
Accepted:2023.08.26
Published online: **2023.09.23**

ABSTRACT

Objective: One of the features of the anchor test, which can affect the equating process, is its correlation with the total test. This systematic review addressed the effects of this feature on the equating process and the factors affecting it.

Methods: To this end, the terms *equating*, *anchor*, *correlation*, and a combination of them were searched on PubMed, Medline, ERIC, JSTOR, and Wiley databases, SAGE, ETS, and ACADEMIA websites, and references of some important articles. The search was restricted to English sources from 1950 to 2022.

Results: Based on the inclusion criteria, 18 out of the 167 extracted documents were selected for further analysis. The quality of documents was measured using the Quality Assessment Tool for Studies with Diverse Designs (QATSDD). The results showed that the test length, test reliability, statistical characteristics of the anchor, the content structure of the anchor test, and differences in the ability of examinee groups were the most important factors affecting the correlation between the anchor test and the total test. The results also demonstrated that the increased correlation between these two tests improved the quality and accuracy of parameter estimation in the equating process and reduced the standard error of equating.

Conclusion: Considering the importance of the correlation between the anchor test and the total test, it is necessary to carefully examine and analyze the value of this correlation and the factors affecting it in the test development process before equating related analysis to minimize errors and biased results.

Keywords: Equating, Anchor Test, Correlation, Systematic Review

Cite this article: Asadi, Vahideh; Moghadamzadeh, Ali; Salehi, Keyvan (2023). The effect of the anchor to total test correlation on equating results: A systematic review. *Educational Measurement and Evaluation Studies*, 13 (43):7-27 pages. DOI: 10.22034/EMES.2023.1971260.2430

Publisher: National Organization of Educational Testing (NOET)

مطالعات اندازه‌گیری و ارزشیابی آموزشی

سازمان سنجش آموزش کشور

شاپا چاپی: ۲۸۶۵-۲۴۷۶ شاپا الکترونیکی: ۰۹۴۲-۲۷۸۳

اثر همبستگی آزمون لنگر با آزمون کل بر نتایج همترازسازی: مرور سیستماتیک

وحیده اسدی[۱]، علی مقدم زاده[۲]، کیوان صالحی[۳]

۱. دانشجوی دکتری سنجش و اندازه‌گیری، دانشکده روانشناسی و علوم تربیتی، دانشگاه تهران، تهران، ایران؛ (نویسنده مسئول)، رایانامه: vahideh.asadi@ut.ac.ir

۲. دانشیار بخش تخصصی پژوهش و سنجش، دانشکده روانشناسی و علوم تربیتی، دانشگاه تهران، تهران، ایران. رایانامه: amoghadamzadeh@ut.ac.ir

۳. دانشیار بخش تخصصی پژوهش و سنجش، دانشکده روانشناسی و علوم تربیتی، دانشگاه تهران، تهران، ایران. رایانامه: keyvansalehi@ut.ac.ir

اطلاعات مقاله

نوع مقاله:
مقاله پژوهشی

دریافت: ۱۴۰۲/۰۱/۰۷
اصلاح: ۱۴۰۲/۰۴/۰۵
پذیرش: ۱۴۰۲/۰۶/۰۴
انتشار: ۱۴۰۲/۰۷/۰۱

چکیده

هدف: یکی از ویژگی‌های آزمون لنگر که از مؤلفه‌های مهم همترازسازی است، همبستگی آن با آزمون کل است. در این مرور سیستماتیک، اثر این ویژگی بر فرایند همترازسازی و عوامل مؤثر بر آن بررسی گردید.

روش پژوهش: یک مرور سیستماتیک بر اساس اطلاعات موجود در پایگاه‌های داده PubMed، Medline، ERIC، JSTOR و Wiley، وب‌سایت‌های SAGE، ETS، ACADEMIA و نیز بررسی منابع مندرج در برخی مقاله‌های مهم اجرا شد. جستجو در بازه زمانی ۱۹۵۰ تا ۲۰۲۲ تنها برای منابع انگلیسی صورت پذیرفت. اصطلاحات جستجو شامل همترازسازی، لنگر و همبستگی بود که با ترکیب آن‌ها، راهبردهای جستجو به دست آمد.

یافته‌ها: با توجه به ملاک‌های ورود، ۱۸ مطالعه از ۱۶۷ منبع جستجو شده، برای بررسی به این مرور راه یافتند. کیفیت این پژوهش‌ها با استفاده از ابزار سنجش کیفیت مطالعه‌ها با طرح‌های مختلف (QATSDD) مورد ارزیابی قرار گرفت. نتایج مطالعه نشان داد که طول آزمون، پایایی آزمون، نوع لنگر از نظر ویژگی‌های آماری، ساختار محتوایی آزمون لنگر و تفاوت در توانایی گروه‌ها، عواملی هستند که بر همبستگی آزمون لنگر و آزمون کل مؤثر است. علاوه بر این، نتایج حاکی از آن بود که با افزایش این همبستگی، کیفیت و دقت برآورد پارامترها در فرایند همترازسازی بهبود می‌یابد و از خطای استاندارد همترازسازی کاسته می‌شود.

نتیجه‌گیری: به دلیل اهمیت همبستگی میان آزمون لنگر و آزمون کل، لازم است مقدار این همبستگی و عوامل مؤثر بر آن در مراحل ساخت آزمون و قبل از اجرای تحلیل‌های مرتبط با همترازسازی به‌دقت بررسی و تحلیل شود تا از بروز خطای همترازسازی و سوگیری در نتایج کاسته شود.

واژه‌های کلیدی: همترازسازی، آزمون لنگر، همبستگی، مرور سیستماتیک

استناد: اسدی، وحیده؛ مقدم‌زاده، علی؛ صالحی، کیوان (۱۴۰۲). اثر همبستگی آزمون لنگر با آزمون کل بر نتایج همترازسازی: مرور سیستماتیک. *مطالعات اندازه‌گیری و ارزشیابی آموزشی*، ۱۳ (۴۳)، ۷-۲۷ صفحه.

DOI:10.22034/EMES.2023.1971260.2430

ناشر: سازمان سنجش آموزش کشور

مقدمه

هنگامی‌که ویرایش‌های جدید یا فرم‌های جایگزین برای یک آزمون با هدف سنجش آزمودنی‌ها طراحی می‌شود، انتظار بر این است که معنای نمره‌ها در این آزمون‌ها یکسان باشد. از دیدگاه شی و نورسینی[۱] (۱۹۹۵) سه دلیل برای کاربرد چند فرم از یک آزمون وجود دارد. اولین دلیل به حفظ امنیت آزمون مرتبط است که بر اساس آن تصمیم‌های سرنوشت‌سازی در حیطه‌های آموزشی و شغلی اتخاذ می‌شود. دلیل دیگر، امکان انتشار سؤال‌های آزمون است. دلیل سوم، تغییر در محتوا و سؤال‌های آزمون در طول زمان است. در چنین شرایطی، اطمینان از یکسان بودن معنای نمره‌ها در فرم‌های مختلف یک آزمون بسیار مهم و ضروری است تا نتایج به‌دست‌آمده برای همه آزمودنی‌ها منصفانه باشد. به همین منظور، طراحان آزمون یک برنامه مشترک برای ساخت آزمون‌هایی با سازه اندازه‌گیری یکسان در نظر می‌گیرند. بااین‌حال، آزمون‌های طراحی‌شده دارای ویژگی‌های روان‌سنجی متفاوتی هستند. برای مقایسه نمره‌ها در این فرم‌ها لازم است میان توابع برآورد نمره‌ها، پیوند برقرار شود تا اثر تفاوت در فرم‌ها، به‌ویژه اثر تفاوت در دشواری حذف گردد. برای دستیابی به این هدف، می‌توان از فرایند همترازسازی[۲] که بین نمره‌های فرم‌های یک آزمون پیوند برقرار می‌کند، استفاده نمود. همترازسازی موجب می‌گردد تا نمره‌های فرم‌های مختلف یک آزمون با سطح دشواری متفاوت، تعدیل شوند (کولن و برنان[۳]، ۲۰۱۴). به دلیل این فرایند، برای آزمودنی فرقی نمی‌کند که به کدام فرم پاسخ داده است؛ بنابراین، همترازسازی، تفاوت در دشواری فرم‌ها را تعدیل می‌کند و هنگامی‌که فرم‌های آزمون با موفقیت هم‌تراز شوند، تفاوت در نمره‌های آزمودنی‌ها به دلیل سختی یا آسانی آزمون نخواهد بود. دورانز و همکاران[۴] (۲۰۱۰) معتقدند که برای به دست آوردن بهترین و دقیق‌ترین برآورد تفاوت در دشواری فرم‌های آزمون، باید در تمام روش‌های همترازسازی، تفاوت گروه‌ها از نظر توانایی کنترل شود. دو رویکرد متفاوت برای دستیابی به این هدف وجود دارد. یک رویکرد، استفاده از جامعه مشترک آزمودنی‌ها و یا دو نمونه معادل از یک جامعه مشترک است. رویکرد دوم، کاربرد سؤال‌های لنگر[۵] در دو فرم آزمون است که طرحی منعطف‌تر را برای جمع‌آوری داده‌ها و اجرای همترازسازی فراهم می‌آورد. سؤال‌های لنگر، سؤال‌های مشترکی[۶] است که از نظر ویژگی‌های محتوایی و آماری مشابه آزمون اصلی است و از نمره‌های آن برای اندازه‌گیری و کنترل تفاوت‌های گروهی استفاده می‌شود (لیوینگستون[۷]، ۲۰۰۴؛ گنزالس و ویبرگ[۸]، ۲۰۱۷). با توجه به این تعریف، اصطلاح سؤال‌های لنگر و سؤال‌های مشترک را می‌توان به‌جای یکدیگر استفاده نمود (رایان و براکمن[۹]، ۲۰۱۸) که اصطلاح سؤال‌های لنگر در مقاله حاضر به کار گرفته شد. دورانز و همکاران (۲۰۰۸) دریافتند نتایج همترازسازی زمانی قابل‌اطمینان است که محتوای آزمون لنگر با آزمون کل مطابقت داشته باشد. هابرمن[۱۰] و دورانز (۲۰۰۹) و وی[۱۱] (۲۰۱۰) گزارش نمودند که آزمون لنگری که از آزمون کل محتوای متفاوتی دارد و مفروضه نمایندگی محتوایی آن نقض شده است، منجر به رانش[۱۲] (انحراف) مقیاس می‌شود. در کنار ویژگی‌های محتوایی، ویژگی‌های آماری آزمون لنگر نیز مهم است. به‌طوری‌که آزمون لنگر باید دارای ویژگی‌های آماری مشابهی نسبت به آزمون کل باشد. کولن و برنان (۲۰۱۴) اشاره نمودند که ویژگی‌های آماری، اغلب بر آماره‌های کلاسیک ازجمله میانگین، انحراف استاندارد، توزیع دشواری و تشخیص سؤال‌ها مبتنی است. هابرمن و دورانز (۲۰۰۹) و وی (۲۰۱۰) اظهار نمودند که هرگاه نمایندگی آماری آزمون لنگر نقض شود، رانش مقیاس رخ خواهد داد. وی (۲۰۱۰) بر این باور است که از بین نمایندگی محتوایی و آماری آزمون لنگر، نمایندگی محتوایی تأثیر بیشتری بر نتایج همترازسازی دارد. علاوه بر موارد ذکرشده، هابرمن و دورانز (۲۰۰۹) بیان نمودند که انتخاب طرح نامناسب، گروه‌های دارای توانایی متفاوت، همبستگی ضعیف بین آزمون لنگر و آزمون کل و آزمون لنگر نامناسب می‌توانند منابع رانش مقیاس باشند. بررسی

---

1. Shea & Norcini
2. equating
3. Kolen & Brennan
4. Dorans et al.
5. anchor items
6. common
7. Livingston
8. Gonzalez & Wiberg
9. Ryan & Brockmann
10. Haberman
11. Wei
12. drift

پیشینه پژوهش‌ها نشان داد که آزمون لنگر، نسخه‌ای کوتاه[1] از آزمون اصلی است (کولن و برنان، ۲۰۱۴، ص. ۱۸) و در بسیاری از برنامه‌های سنجش که از طرح گروه‌های نامعادل با آزمون لنگر[2] (NEAT) (ون‌داویر و همکاران[3]، ۲۰۰۴) استفاده می‌کنند، به‌عنوان یک آزمون کوچک (minitest) شناخته و در نظر گرفته می‌شود (سینهارای و هالند، ۲۰۰۶b، ص. ۱؛ لیو و همکاران[4]، ۲۰۱۱a؛ سینهارای، ۲۰۱۷). نتایج پژوهش‌های سینهارای و هالند (۲۰۰۶b، ۲۰۰۷) و لیو و همکاران (۲۰۱۱a، ۲۰۱۱b) در ارتباط با نوع آزمون لنگر نشان داد که سؤال‌های آزمون لنگر با تغییرپذیری کم در دشواری، نتایج همترازسازی باثباتی ایجاد می‌کند. آزمون لنگر در همترازسازی نمره‌های آزمون نقش کلیدی دارد و انتخاب آن در کیفیت همترازسازی، به‌ویژه هنگام استفاده از طرح NEAT بسیار بااهمیت است (سینهارای و همکاران، ۲۰۱۲؛ سینهارای، ۲۰۱۷). یکی از عوامل تعیین‌کننده کارایی فرایند همترازسازی، همبستگی[5] بین آزمون لنگر و آزمون کل است (بیدسکو[6]، ۱۹۸۵). سؤالی که مطرح می‌شود این است که همبستگی آزمون لنگر و آزمون کل چه تأثیری بر فرایند همترازسازی دارد؟ تاکنون هیچ پژوهشی به‌طور سیستماتیک بر اساس مطالعه‌های موجود برای پاسخ به این سؤال انجام‌نشده است. با توجه به دانش موجود در زمینهٔ ویژگی‌های آزمون لنگر، این مطالعه به‌منظور انجام یک مرور سیستماتیک برای بررسی اثر همبستگی آزمون لنگر با آزمون کل بر نتایج همترازسازی طراحی شد. اهمیت این مطالعه ازآن‌جهت است که نتایج پژوهش‌های انجام‌شده در مورد اثر همبستگی با جزئیات بیشتری موردبررسی قرار می‌گیرد و از ترکیب این نتایج مشخص می‌شود که وجود همبستگی بالا بین این دو آزمون چه تأثیری بر خطای همترازسازی دارد. همچنین از این مطالعه می‌توان به عواملی که بر میزان همبستگی این آزمون‌ها اثر می‌گذارد، پی برد. بر اساس نتایج این مرور، طراحان با شناخت و درک عوامل مؤثر بر همبستگی آزمون لنگر و آزمون کل می‌توانند در هنگام ساخت فرم‌های مختلف از یک آزمون، آن‌ها را در نظر بگیرند تا نتایج همترازسازی بهبود یابد و حداقل خطا رخ دهد.

روش‌ها

پروتکل

قبل از نوشتن مقاله، پروتکلی برای انجام این مرور سیستماتیک طراحی شد. در نگارش پروتکل، علاوه بر ارائه مقدمه‌ای در مورد چرایی انتخاب موضوع و اهمیت اجرای آن، هدف‌های پژوهش، سؤال‌های مرور سیستماتیک، ملاک‌های ورود منابع برای اجرای پژوهش، راهبردهای پژوهش، نحوه انتخاب مطالعه‌ها، فرم‌های ارزیابی انتقادی و سنجش کیفیت مطالعه‌ها، نحوه استخراج داده‌ها و روش ترکیب داده‌های به‌دست‌آمده به‌تفصیل در نظر گرفته شد. راهبردهای جستجو برای هر پایگاه داده و وب‌سایت در یک فایل جداگانه طراحی و نوشته شد. درنهایت برای ارزیابی پروتکل طراحی‌شده از فهرست‌وارسی[7] پروتکل پریزما[8] سال ۲۰۱۵ (شمسیر و همکاران[9]، ۲۰۱۵) استفاده شد.

هدف‌ها

هنگام همترازسازی آزمون‌ها با طرح NEAT، عموماً این باور وجود دارد که آزمون لنگر باید نسخه کوچک آزمون‌هایی باشد که هم‌تراز می‌شوند (لیو و همکاران، ۲۰۱۱a؛ کولن و برنان، ۲۰۱۴؛ سینهارای، ۲۰۱۷). در این طرح که یکی از منعطف‌ترین ابزارهای موجود برای همترازسازی آزمون‌ها است (انگاف[10]، ۱۹۷۱؛ پترسن و همکاران[11]، ۱۹۸۲؛ پترسن و همکاران، ۱۹۸۹؛ کولن و برنان، ۲۰۰۴؛ سینهارای و هالند، ۲۰۰۶b)، محاسبه تفاوت‌های گروهی مبتنی بر کاربرد یک آزمون لنگر است (بران[12] و هالند، ۱۹۸۲؛ کولن و برنان، ۲۰۰۴؛ ون‌داویر و همکاران، ۲۰۰۴؛ ماسز[13] و

1. mini-version
2. non-equivalent groups with anchor test (NEAT)
3. von Davier et al.
4. Liu et al.
5. correlation
6. Budescu
7. checklist
8. preferred reporting items for systematic review and meta-analyses (PRISMA)
9. Shamseer et al.
10. Anogff
11. Petersen et al.
12. Braun
13. Moses et al.

همکاران، ۲۰۱۰). در طرح NEAT، داده‌ها از دو جامعه نامعادل (Q, P) که دو آزمون مختلف (Y, X) و یک آزمون لنگر (A) دریافت کردند، جمع‌آوری می‌شود (سینهارای و هالند، ۲۰۰۶b؛ ماسز و همکاران، ۲۰۱۰). هدف این همترازسازی، ایجاد یک تبدیل از نمره X به نمره Y است که تفاوت دشواری فرم‌های آزمون را حذف می‌کند (ماسز و همکاران، ۲۰۱۰). هدف از این مطالعه، انجام یک مرور سیستماتیک برای تعیین این‌که چگونه همبستگی آزمون لنگر و آزمون کل بر نتایج همترازسازی تأثیر می‌گذارد و چه عواملی بر این همبستگی مؤثر هستند، است؛ بنابراین، این پژوهش دو هدف ویژه دارد. هدف اول، تعیین عوامل مؤثر بر همبستگی این دو آزمون و هدف دوم، بررسی اثر این همبستگی بر نتایج همترازسازی است.

سؤال‌های مرور

یکی از مراحل انجام یک مرور سیستماتیک، طراحی سؤال برای آن است. بررسی منابع مرتبط با مرور سیستماتیک نشان داد که برای طرح این سؤال‌ها باید چند مؤلفه را در نظر گرفت تا چرایی انجام مرور توضیح داده شود و واضح‌تر بیان گردد. یکی از اصطلاحاتی که این مؤلفه‌ها را به‌خوبی معرفی می‌کند، PICO است (سانتوس و همکاران[۱]، ۲۰۰۷؛ لاسرسون و همکاران[۲]، ۲۰۱۹؛ تای و همکاران[۳]، ۲۰۲۰) که شامل چهار مؤلفهٔ مسئله[۴] (جامعه)[۵] (P)، مداخله[۶] (I)، مقایسه[۷] (C) و پیامد[۸] (O) است. بر این اساس بخش‌های اصلی سؤال این مرور را می‌توان با اصطلاح PICO به‌صورت زیر بیان کرد: مسئله (P) موردبررسی، همبستگی است. آنچه در معرض مداخله (I) قرار دارد، ویژگی‌های آزمون لنگر است. در بخش مقایسه (C)، همبستگی آزمون لنگر از نظر نوع، طول آزمون و پایایی با آزمون اصلی موردنظر است. اثر همبستگی آزمون لنگر و آزمون کل بر فرایند همترازسازی و هم‌چنین تعیین عوامل مؤثر بر همبستگی این دو آزمون، پیامد (O) این مطالعه است؛ بنابراین، سؤال اصلی پژوهش این است که کدام ویژگی‌های آزمون لنگر بر این همبستگی مؤثر است و این همبستگی چه تأثیری بر فرایند همترازسازی دارد؟ درنهایت دو سؤال این مرور به شرح زیر است.

۱. کدام عوامل بر همبستگی بین آزمون لنگر و آزمون کل مؤثر هستند؟
۲. همبستگی آزمون لنگر و آزمون کل چه اثری بر نتایج همترازسازی دارد؟

ملاک‌های ورودی

ملاک‌های موردنظر برای ورود مطالعه‌ها بر اساس طرح ارائه‌شده، به شرح زیر است: (۱) مطالعه‌ای دارای شایستگی است که یافته‌های تجربی در مورد اثر همبستگی آزمون لنگر با آزمون کل بر فرایند همترازسازی را بیان نماید. (۲) مطالعه‌ای که یافته‌های تجربی را در مورد عوامل مؤثر بر همبستگی آزمون لنگر با آزمون کل گزارش می‌کند، شایسته انتخاب است. (۳) ازآنجایی‌که این مطالعه بر آزمون لنگر متمرکز است، تنها مطالعه‌هایی انتخاب می‌شوند که تحت طرح همترازسازی NEAT انجام‌شده باشند. (۴) مطالعه‌های منتخب، می‌توانند آزمون لنگر درونی و یا لنگر بیرونی داشته باشند. (۵) پژوهش‌هایی که انتخاب می‌شوند، از نظر نوع داده ممکن است دارای داده‌های واقعی[۹] و یا شبیه‌سازی[۱۰] باشند. (۶) از نظر روش تحلیل داده‌ها، مطالعه‌هایی با روش نظریه کلاسیک[۱۱] (CTT)، نظریه سؤال ـ پاسخ[۱۲] (IRT)، مدل راش[۱۳] و همترازسازی کرنل[۱۴] (KE) را

1. Santos et al.
2. Lasserson et al.
3. Tai et al.
4. problem
5. population
6. intervention
7. comparison
8. outcome
9. real data
10. simulation
11. classical test theory (CTT)
12. item response theory (IRT)
13. Rasch model
14. kernel equating (KE)

می‌توان برای این مرور انتخاب کرد. (۷) پژوهش‌های منتخب می‌توانند تحلیلی از نوع اولیه و یا ثانویه[1] داشته باشند. (۸) مطالعه‌های انتخاب‌شده باید دارای یک آزمون چندگزینه‌ای با سؤال‌هایی از نوع دو ارزشی باشند. علاوه بر موارد ذکرشده برای ورود مطالعه‌ها به این مرور، محدودیتی در مورد جامعه تحت بررسی پژوهش‌ها در نظر گرفته نشد. مقاله‌ها، پایان‌نامه‌ها، رساله‌ها و مطالب منتشرشده در کنفرانس‌های معتبر که دارای ملاک‌های موردنظر بوده و تمام متن آن در دسترس بود، انتخاب شدند و مورد تحلیل قرار گرفتند.

راهبردهای جستجو

همان‌طور که اشاره شد، هنگام طراحی یک پروتکل برای راهبردهای جستجو، ابتدا اصطلاحاتی بر اساس عنوان و هدف‌های پژوهش در نظر گرفته شد. پس‌ازآن، هر یک از این اصطلاحات جستجو[2] با مترادف‌های خود با استفاده از عملگرهای بولی[3] (AND/OR) برای تعریف مسیر جستجو ترکیب شدند. اصطلاحات جستجوی اولیه شامل "Equating"، "Anchor"، "Correlation"، 'NEAT'، "Effect" و "Test" بود. علاوه بر جستجو در تمام زمینه‌ها[4]، فیلترهای جستجو شامل عنوان، چکیده و واژه‌های کلیدی بود. این اطلاعات برای جستجو به‌صورت جدولی برای هر پایگاه داده و وب‌سایت موردنظر طراحی شد. پایگاه‌های اطلاعاتی قابل‌دسترس برای پژوهشگر و اجرای این مرور سیستماتیک شامل PubMed، Medline، ERIC، JSTOR و Wiley بود. این جستجوها در بازه زمانی ۱۹۵۰ تا ۲۹ می ۲۰۲۲ (۱۴۰۱/۳/۸) صورت پذیرفت. ازآنجایی‌که تاکنون پژوهشی در این زمینه به زبان فارسی انجام‌نشده است، تمام جستجوها صرفاً در زبان انگلیسی انجام گرفت. علاوه بر جستجو در سطح پایگاه داده، برخی از وب‌سایت‌ها مانند SAGE، ETS و ACADEMIA برای بررسی بیشتر جستجو شدند. برای برخی از مجله‌ها و کتاب‌های تخصصی، جستجوهای دستی انجام شد. منابع مقاله‌های مهم نیز برای مطالعه بیشتر و دقت در جستجوی منابع مرتبط بررسی شدند. این اقدامات به‌منظور جستجوی کلیه اطلاعات مرتبط با موضوع این مرور صورت پذیرفت تا یک بررسی جامع و چندجانبه انجام شود و نتایج کامل‌تری به دست آید. جدول ۱ برخی از اطلاعات مربوط به این جستجوها را ارائه می‌دهد.

جدول ۱. خلاصه‌ای از راهبردهای جستجو

نوع	جستجو	اصطلاحات جستجو	تعداد
پایگاه داده	PubMed	Equating, Anchor, Correlation	۱۰
	Medline	Equating, Anchor, Correlation, NEAT, Effect, Test	۵۰
	ERIC	Equating, Anchor, Correlation, NEAT, Effect, Test	۲۱
	JSTOR	Equating, Anchor, Correlation, NEAT, Effect	۲٤
	Wiley	Equating, Anchor, Correlation	۱۱
وب‌سایت	SAGE	Equating, Anchor, Correlation, NEAT	۱۳
	ETS	Equating, Anchor, Correlation	٦
	ACADEMIA	Equating, Anchor, Correlation	۱۹
موارد دیگر			۱۳

زمان: از سال ۱۹۵۰ تا ۲۹ می ۲۰۲۲ (۱٤۰۱/۳/۸)

فیلترها در سطح: همه حوزه‌ها، عنوان، چکیده و کلیدواژه‌ها

زبان: انگلیسی

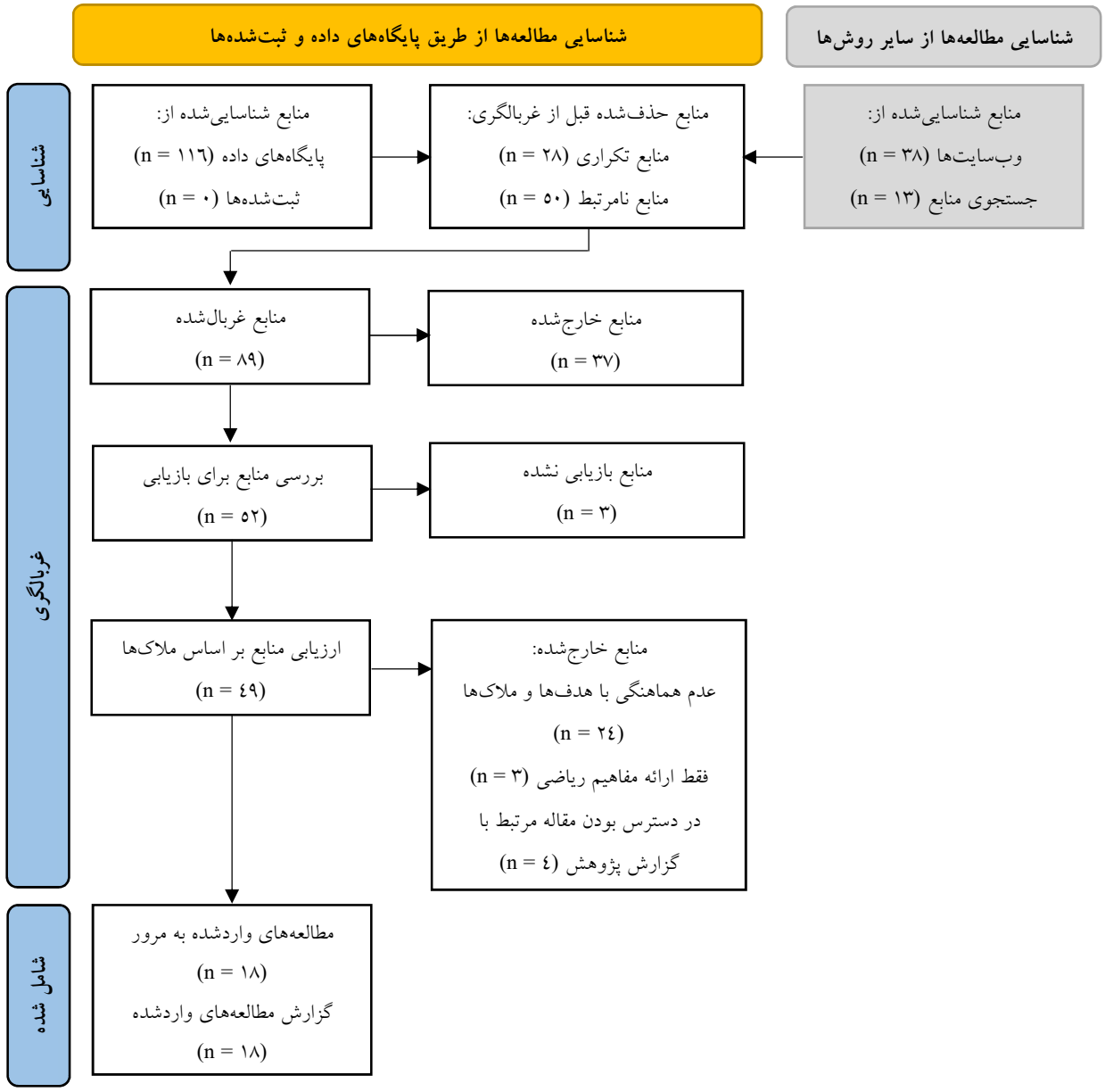
[1]. secondary analysis
[2]. search terms
[3]. boolean operator
[4]. all fields

در جدول بالا ملاحظه می‌شود که تعداد کل منابع یافت شده بر اساس مسیرهای جستجو ۱۶۷ مورد است. پس از اتمام جستجو، تمام این منابع، به نرم‌افزار EndNote (نسخه ۲۰. ۵) برای حذف مطالعات تکراری[1]، منابع بی‌ارتباط و نامرتبط منتقل شد.

انتخاب مطالعه

همه مطالعه‌ها (۱۶۷ مورد) به نرم‌افزار EndNote منتقل شدند تا منابع مناسب، انتخاب شوند. پس از شناسایی و حذف منابع تکراری (۲۸ مورد)، منابعی که عنوان آن‌ها کاملاً نامرتبط به موضوع مرور بودند (۵۰ مورد) از منابع خارج شدند. در بین ۸۹ منبع باقی‌مانده، منابعی که عنوان آن‌ها با موضوع پژوهش حاضر مطابق بود، حفظ شدند. از منابع نهایی (۵۲ مورد)، ۳ منبع دانلود نشد. ۴۹ منبع باقی‌مانده که فایل آن‌ها به‌طور کامل در دسترس بود، با توجه به اهداف مرور بر اساس عنوان، چکیده و متن بررسی شدند و درنهایت ۲۴ منبع به دلیل نداشتن ملاک‌های اولیه حذف شدند. با توجه به ملاک‌های ورود، ۲۵ منبع باقی‌مانده به‌طور جداگانه مورد تحلیل قرار گرفت که منجر به حذف ۷ منبع شد. مقاله‌های ون‌داویر و همکاران (۲۰۰۴)، ون‌داویر (۲۰۰۸) و هابرمن و دورانز (۲۰۰۹) تنها به موضوع پژوهش خود از منظر نظری پرداختند و ازآنجایی‌که امکان سنجش کیفیت این مقاله‌ها فراهم نبود، این سه مقاله حذف شدند؛ اما از نتایج آن‌ها در پاسخ به سؤال‌های پژوهش استفاده شد. مطالعه‌های سینهارای و هالند (۲۰۰۶b)، لیو و همکاران (۲۰۰۹)، پوهان[2] (۲۰۱۰) و سینهارای و همکاران (۲۰۱۲)، گزارش‌های پژوهشی منتشرشده توسط شرکت خدمات سنجش آموزشی[3] (ETS) هستند. جستجوی بیشتر نشان داد که مقاله‌هایی بر اساس این مطالعه‌ها ارائه‌شده است. ازآنجایی‌که محتوای این گزارش‌ها مشابه مقاله‌های منتشرشده بود، این ۴ منبع حذف شدند. با توجه به منابع حذف‌شده، ۱۸ منبع برای ورود به این مرور و انجام بررسی‌های بیشتر باقی ماند. انتخاب مطالعه‌ها طی این مراحل با استفاده از نمودار پریزما (۲۰۲۰) (پیج و همکاران[4]، ۲۰۲۱) در شکل ۱ گزارش‌شده است.

1. duplicate
2. Puhan
3. Educational Testing Service (ETS)
4. Page et al.



شکل ۱. نمودار پریزما (۲۰۲۰) برای انتخاب مطالعه‌ها

سنجش کیفیت مطالعه‌ها

برای ادامه روند مرور و قبل از ارائه نتایج، گام مهم، ارزیابی کیفیت روش‌شناختی منابع انتخاب‌شده است که نیازمند انتخاب ابزار مناسب برای سنجش کیفیت مطالعه‌های موردنظر است. بررسی کیفیت پژوهش‌های انجام‌شده به ایجاد اعتماد در نتایج این مطالعه‌ها کمک می‌کند. انواع مختلفی از ابزارهای سنجش کیفیت[1] (QA) (مانند CASP، COSMIN، CONSORT و CEBM) و خطر سوگیری[2] (ROB) (مثل ROBINS و ACROBAT-NRSI) (برای مطالعه بیشتر به پیچ و همکاران (۲۰۱۸) مراجعه شود) توسعه‌یافته‌اند. برخی از پژوهشگران این دو اصطلاح را

[1]. quality assessment (QA)
[2]. risk of bias assessment (ROB)

معادل هم می‌دانند و به‌جای یکدیگر بکار می‌برند. درحالی‌که کانامورى و همکاران[1] (۲۰۲۱) در مقاله‌ای بیان نمودند که این دو اصطلاح متفاوت‌اند و قابل‌معاوضه نیستند. علاوه بر معنای متفاوت این دو اصطلاح، ابزارهای متفاوتی برای آن‌ها طراحی‌شده است. ازآنجایی‌که ابزارهای سنجش خطر سوگیری اغلب برای کارآزمایی‌های بالینی، مطالعه‌های مداخله‌ای و مشاهده‌ای طراحی می‌شوند، برای این مرور که شامل مطالعه‌های کمّی، همبستگی و توصیفی است، فاقد تناسب و کاربرد است؛ بنابراین، تنها سنجش کیفیت مطالعه‌های انتخاب‌شده، مدنظر قرار گرفت. یکی از ابزارهایی که می‌تواند برای سنجش کیفیت مطالعه‌های کمّی مورداستفاده قرار گیرد، ابزار سنجش کیفیت مطالعه‌ها با طرح‌های مختلف[2] (QATSDD) (فنتون و همکاران[3]، ۲۰۱۵) است. این ابزار دارای ۱۶ نشانگر[4] بر اساس مقیاس لیکرت ۴ درجه‌ای (از ۰ تا ۳) است. قالب این ابزار دارای فرمی است که می‌توان از آن برای پژوهش‌های کمّی، کیفی و ترکیبی استفاده کرد (سیریه و همکاران[5]، ۲۰۱۱). این نشانگرها شامل، چارچوب نظری روشن، بیان هدف‌ها در متن اصلی گزارش، توصیف واضح زمینه پژوهش، شواهدی از اندازه نمونه در نظر گرفته‌شده، نماینده بودن نمونه گروه هدف، توصیف جمع‌آوری داده‌ها، دلیل انتخاب ابزار جمع‌آوری داده‌ها، اطلاعات دقیق بکار گرفته‌شده، ارزیابی آماری پایایی و روایی ابزار اندازه‌گیری، تناسب سؤال‌های پژوهش و روش جمع‌آوری داده‌ها، تناسب سؤال‌های پژوهش و محتوای ابزار جمع‌آوری داده‌ها، تناسب سؤال‌های پژوهش و روش تحلیل، توجیه روش‌های تحلیل انتخاب‌شده، ارزیابی پایایی فرایند تحلیل، شواهد مشارکت کاربران در طرح و بحث انتقادی درباره نقاط قوت و محدودیت‌ها است. از بین این نشانگرها، موارد ۹ و ۱۰ فقط برای پژوهش‌های کمّی و موارد ۱۱ و ۱۴ تنها برای پژوهش‌های کیفی مورداستفاده قرار می‌گیرند. با توجه به نوع مطالعه‌های این مرور که از نوع کمّی هستند، موارد ۱۱ و ۱۴ قابل‌اجرا نیستند[6] و در جدول NA آمده است. ضمناً، مورد ۱۵ نیز برای این مطالعه‌ها قابل‌اجرا نبود و موردبررسی قرار نگرفت. با توجه به آنچه بیان شد، برای هر مطالعه بر اساس ۱۳ نشانگر باقی‌مانده، بررسی‌هایی انجام گرفت. نتایج سنجش کیفیت برای هر یک از مطالعه‌ها در جدول ۲ ارائه‌شده است. همان‌طور که در جدول ملاحظه می‌شود، همه مطالعه‌ها دارای نمرهٔ بالایی هستند (بالاتر از ۷۰ درصد) که نشان‌دهنده کیفیت مطلوب این مطالعه‌ها است.

جدول ۲. سنجش کیفیت مطالعه‌های منتخب با QATSDD و نمره‌گذاری آن‌ها

		نشانگرها																				
منبع	سال	۱	۲	۳	٤	٥	٦	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	۱٤	۱٥	۱٦	نمره	درصد
بیدسکو	۱۹۸۵	۳	۲	۲	۰	۰	۳	۳	۳	۳	۳	NA	۳	۳	NA	NA	۲	۳۰	۷٦/۹۲
بالا	۱۹۸۸	۳	۳	۳	۳	۳	۳	۳	۳	۰	۳	NA	۳	۳	NA	NA	۲	۳٥	۸۹/۷٤
یانگ و هوانگ	۱۹۹٦	۳	۳	۳	۳	۳	۳	۲	۳	۱	۳	NA	۳	۳	NA	NA	۲	۳٥	۸۹/۷٤
سینهارای و هالند	۲۰۰٦a	۳	۳	۳	۳	۳	۳	۳	۳	۱	۳	NA	۳	۳	NA	NA	۲	۳٦	۹۲/۳۰
سینهارای و هالند	۲۰۰۷	۳	۲	۲	۳	۳	۳	۳	۳	۰	۳	NA	۳	۳	NA	NA	۳	۳٤	۸۷/۱۷
ماسز و کیم	۲۰۰۷	۳	۲	۲	۳	۳	۳	۳	۳	۳	۳	NA	۳	۳	NA	NA	۲	۳٦	۹۲/۳۰
ریکر و ون‌داویر	۲۰۰۷	۳	۲	۲	۱	۳	۳	۱	۳	۳	۳	NA	۳	۳	NA	NA	۰	۳۰	۷٦/۹۲
سو و همکاران	۲۰۰۹	۳	۱	۱	۳	۳	۳	۳	۳	۳	۳	NA	۳	۱	NA	NA	۰	۳۰	۷٦/۹۲
یی	۲۰۰۹	۳	۳	۳	۲	۳	۳	۳	۲	۰	۳	NA	۳	۳	NA	NA	۲	۳۳	۸٤/٦۱
پوهان	۲۰۱۰	۳	۳	۳	۳	۳	۳	۲	۳	۳	۳	NA	۳	۳	NA	NA	۳	۳۸	۹۷/٤۳
سان‌ناسی	۲۰۱۱	۳	۳	۳	۳	۳	۳	۳	۳	۳	۳	NA	۲	۲	NA	NA	۳	۳۷	۹٤/۸۷
لیو و همکاران	۲۰۱۱a	۳	۲	۳	۱	۳	۳	۳	۳	۳	۳	NA	۳	۳	NA	NA	۲	۳٥	۸۹/۷٤

[1]. Kanamori et al.
[2]. quality assessment tool for studies with diverse design (QATSDD)
[3]. Fenton et al.
[4]. indicator
[5]. Sirriyeh et al.
[6]. not applicable (NA)

لیو و همکاران	۲۰۱۱b	۳	۳	۲	۱	۳	۳	۳	۳	۳	NA	۳	۳	NA	NA	۳	۳۶	۹۲/۳۰
ژانگ و کولن	۲۰۱۳	۳	۳	۲	۳	۳	۳	۳	۳	۳	NA	۳	۳	NA	NA	۲	۳۷	۹٤/۸۷
تری‌ویلر و همکاران	۲۰۱٦	۳	۳	۳	۱	۳	۳	۳	۳	۳	NA	۳	۳	NA	NA	۲	۳٦	۹۲/۳۰
لین و همکاران	۲۰۱٦	۲	۳	۳	۱	۳	۳	۳	۲	۰	NA	۳	۳	NA	NA	۱	۳۰	۷٦/۹۲
سینهارای	۲۰۱۷	۳	۳	۳	۱	۰	۳	۳	۳	NA	NA	۳	۳	NA	NA	۲	۲۹	۸۰/۵۵
مارنگو و همکاران	۲۰۱۸	۳	۳	۲	۲	۳	۳	۳	۳	۳	NA	۳	۳	NA	NA	۳	۳۷	۹٤/۸۷

قابل‌اجرا نیست (not applicable, NA)

نتایج

در میان تمام مطالعه‌هایی که در طول فرایند جستجوی این مرور یافت شد، ۱۸ منبع که ملاک‌های ورود به مرور را داشتند، دقیق‌تر موردبررسی قرار گرفتند. جزئیات این اطلاعات در جدول ۳ ارائه‌شده است. همه این مطالعه‌ها، تابع همترازسازی را تحت طرح گروه‌های نامعادل با آزمون لنگر (NEAT) برای داده‌های دو ارزشی بکار برده و تحلیل نموده‌اند. نحوه قرار گرفتن سؤال‌های آزمون لنگر در هر دو آزمون می‌تواند درونی[1] یا بیرونی[2] باشد (پترسن، ۲۰۰۷؛ دورانز و همکاران، ۲۰۱۰، ۲۰۱۱؛ رایان و براکمن، ۲۰۱۸). آزمون لنگر بکار رفته در مطالعه‌های این مرور، از نظر جایگاه و موقعیت، شامل آزمون لنگر درونی (۵ مورد، ٪ ۲۷/۷۷)، بیرونی (۹ مورد، ٪ ۵۰) و یا هر دو نوع (۴ مورد، ٪ ۲۲/۲۲) بود. از نظر ویژگی‌های آماری (میانگین و واریانس دشواری سؤال‌ها)، سه نوع لنگر mini، midi و semi-midi معرفی‌شده است (سینهارای و هالند، ۲۰۰۶b). لازم به ذکر است که در این مقاله به‌منظور حفظ مفهوم اولیه این سه نوع لنگر و سهولت استفاده از آن‌ها در هنگام گزارش نتایج، نام آن‌ها بدون ترجمه و بر اساس همان اصطلاحات اصلی در متن پژوهش ذکرشده است. در این مطالعه‌ها، آزمون لنگر بر اساس ویژگی‌های آماری از نوع mini (۱۱ مورد، ٪ ۶۱/۱۱)، mini و midi (۳ مورد، ٪ ۱۶/۶۶) و یا هر سه نوع (۴ مورد، ٪ ۲۲/۲۲) بود. کولن (۲۰۲۰) اظهار داشت که استفاده از داده‌های واقعی دارای مزایایی است که ازجمله آن می‌توان به کنترل خطای تصادفی و واقع‌گرایی مطالعه و نتایج آن اشاره کرد. از میان این مطالعه‌ها، ۵ مورد (٪۲۷/۷۷) از داده‌های واقعی، ۱۰ مورد (٪ ۵۵/۵۵) از داده‌های شبیه‌سازی‌شده و ۳ مورد (٪ ۱۶/۶۶) از هر دو نوع داده استفاده کرده بودند. از نظر نوع توزیع داده‌ها، ۱۳ مورد (٪ ۷۲/۲۲) آن‌ها دارای توزیع نرمال، ۱ مورد (٪ ۵/۵۵) توزیع با کجی منفی، ۱ مورد (٪ ۵/۵۵) دارای توزیع شرطی، ۱ مورد (٪ ۵/۵۵) با توزیع حاشیه‌ای و دو مطالعه (٪ ۱۱/۱۱) بدون ذکر نوع توزیع بودند. یانگ و هوانگ[3] (۱۹۹۶) در پژوهش خود تأثیر طول آزمون لنگر بر دقت نتایج روش‌های مختلف همترازسازی را با استفاده از طرح NEAT بررسی کردند. برای انجام این مطالعه از داده‌های واقعی استفاده شد که دارای توزیعی با کجی منفی بود. با توجه به نوع توزیع، در تحلیل‌ها، پارامتر حدس را بکار بردند. نتایج نشان داد که همترازسازی آزمون با در نظر گرفتن پارامتر حدس در توزیعی با کجی منفی از کفایت و تناسب لازم برخوردار است. از دیدگاه لرد[4] (۱۹۷۷)، «اگر و فقط اگر تفاوتی نکند که آزمودنی به فرم X یا Y پاسخ دهد، تبدیل نمره X به Y همترازسازی به‌حساب می‌آید» (ص. ۱۲۸). حال اگر دو فرم آزمون از نظر توانایی، پایایی و دشواری متفاوت باشند، این تعریف نقض می‌شود و نمی‌توان دو فرم آزمون را هم‌تراز کرد. رویکردهای همترازسازی آزمون مبتنی بر CTT، IRT و راش برای ایجاد نمره‌های قابل‌مقایسه در آزمون‌هایی که با حداقل تفاوت در دشواری طراحی‌شده‌اند، استفاده می‌شود. همترازسازی کرنل یک رویکرد یکپارچه برای هم‌تراز کردن دو فرم آزمون است و شامل مجموعه‌ای از توابع همترازسازی مشابه همترازسازی همصدک است (ون‌داویر و همکاران، ۲۰۰۴). این روش از هموارسازی کرنل برای ایجاد پیوستگی توزیع نمره‌های گسسته استفاده می‌کند (والین و همکاران[5]، ۲۰۲۱). روش‌های تحلیل در نظر گرفته‌شده در این مطالعه‌ها شامل CTT (۷ مورد، ٪ ۳۸/۸۸)، IRT (۳ مورد، ٪ ۱۶/۶۶)، راش (۱ مورد، ٪ ۵/۵۵) و بیشتر از یک روش (۶ مورد، ٪ ۳۳/۳۳) بود. همان‌طور که در ملاک‌های ورود به مطالعه ذکر شد، محدودیتی

1. internal
2. external
3. Yang & Houang
4. Lord
5. Wallin et al.

برای جامعه و نمونه‌ها در نظر گرفته نشد تا بتوان طیف وسیع‌تری از مطالعه‌ها و نتایج آن‌ها را بررسی کرد. دامنه نمونه‌های این مطالعه‌ها از ۲۵ تا ۲۰۰۰۰ نفر متغیر بود. برخی از این مطالعه‌ها از داده‌های آزمون‌هایی مانند SAT[1] و MBE[2] استفاده کردند. تعدادی از مطالعه‌ها، داده‌های پژوهش‌های قبلی را بکار بردند، مانند پژوهش تری‌ویلر و همکاران[3] (۲۰۱۶) (داده‌های مطالعه سینهارای و هالند (۲۰۰۶a)) و سینهارای (۲۰۱۷) (بررسی داده‌های پژوهش تری‌ویلر و همکاران (۲۰۱۶) که نام آن را TLS16 گذاشتند). ممکن است این سؤال پیش بیاید که چرا ملاک‌های ورود به این پژوهش، متنوع بوده و تنها بر یک نوع خاص متمرکز نشده است؟ دلیل این مورد این است که مشخص شود آیا پاسخ به سؤال‌های پژوهش در شرایط مختلف، یکسان است یا این‌که نتایج مرور به دلیل تفاوت در ملاک‌های ورود به مطالعه، متفاوت خواهد بود؟

جدول ۳. مشخصات مطالعه‌های واردشده به مرور سیستماتیک

منبع	سال	تابع	طرح	لنگر	لنگر	داده‌ها	سؤال‌ها	توزیع	روش تحلیل	نمونه
بیدسکو	۱۹۸۵	همترازسازی	NEAT	درونی	mini	واقعی	دو ارزشی	نرمال	CTT	SAT
بالا	۱۹۸۸	همترازسازی	NEAT	بیرونی	mini	شبیه‌سازی	دو ارزشی	نرمال	CTT/IRT	۱۰۰۰
یانگ و هوانگ	۱۹۹۶	همترازسازی	NEAT	درونی بیرونی	mini	واقعی	دو ارزشی	کجی منفی	CTT IRT	۲۲۴۱
سینهارای و هالند	۲۰۰۶a	همترازسازی	NEAT	درونی بیرونی	mini midi semi	شبیه‌سازی مثال واقعی	دو ارزشی	نرمال	IRT Rasch	۱۰۰۰ ۲۰۰۰ ۶۴۸۹
سینهارای و هالند	۲۰۰۷	همترازسازی	NEAT	بیرونی	mini midi semi	شبیه‌سازی مثال واقعی	دو ارزشی	نرمال	IRT	۱۰۰ ۵۰۰ ۵۰۰۰
ماسز و کیم	۲۰۰۷	همترازسازی	NEAT	بیرونی	mini	شبیه‌سازی	دو ارزشی	نرمال	CTT CM GT	۵۰۰ ۱۰۰۰ ۵۰۰۰
ریکر و ون‌داویر	۲۰۰۷	همترازسازی	NEAT	بیرونی	mini	شبیه‌سازی	دو ارزشی	نرمال	CTT KE	۴۲۳۷ ۶۱۶۸ ۱۰۴۰۵
سو و همکاران	۲۰۰۹	همترازسازی	NEAT	درونی	mini	شبیه‌سازی	دو ارزشی	نرمال	CTT	MBE ۲۰۰۰۰ ۲۰۰۰۰
یی	۲۰۰۹	همترازسازی	NEAT	درونی	mini midi	شبیه‌سازی مثال واقعی	دو ارزشی	نرمال	IRT	۵۰۰۰
پوهان	۲۰۱۰	همترازسازی	NEAT	درونی	mini	شبیه‌سازی	دو ارزشی	–	CTT	۱۰۰۰
سان‌ناسی	۲۰۱۱	همترازسازی	NEAT	درونی	mini	شبیه‌سازی	دو ارزشی	نرمال	CTT	۲۵ ،۵۰ ۱۰۰،۲۰۰ ۴۰۰

[1]. Scholastic Aptitude Test (SAT)
[2]. Multistate Bar Examination (MBE)
[3]. Trierweiler et al.

لیو و همکاران	۲۰۱۱a	همترازسازی	NEAT	بیرونی	mini midi	واقعی	دو ارزشی	شرطی	CTT IRT	SAT
لیو و همکاران	۲۰۱۱b	همترازسازی	NEAT	بیرونی	mini midi	واقعی	دو ارزشی	حاشیه‌ای	CTT	SAT
ژانگ و کولن	۲۰۱۳	همترازسازی	NEAT	درونی بیرونی	mini	شبیه‌سازی	دو ارزشی	نرمال	CTT	۲۰۰۰
تری‌ویلر و همکاران	۲۰۱۶	همترازسازی	NEAT	بیرونی	mini midi semi	شبیه‌سازی	دو ارزشی	نرمال	IRT	۱۰۰۰
لین و همکاران	۲۰۱۶	همترازسازی	NEAT	بیرونی	mini	شبیه‌سازی	دو ارزشی	نرمال	CTT	–
سینهارای	۲۰۱۷	همترازسازی	NEAT	درونی بیرونی	mini midi semi	شبیه‌سازی	دو ارزشی	–	–	TLS16
مارنگو و همکاران	۲۰۱۸	همترازسازی	NEAT	بیرونی	mini	واقعی	دو ارزشی	نرمال	Rasch	۱۸۱۳

گروه‌های نامعادل با آزمون لنگر (non-equivalent groups with anchor test, NEAT)؛ (minitest) mini؛ (miditest) midi؛ (semi-miditest) semi؛ نظریه کلاسیک آزمون (classical test theory, CTT)؛ نظریه سؤال ـ پاسخ (item-response theory, IRT)؛ مدل راش (Rasch)؛ مدل متجانس (congeneric model, CM)؛ نظریه تعمیم‌پذیری (generalizability theory, GT)؛ همترازسازی کرنل (kernel equating, KE)؛ آزمون استعداد تحصیلی (Scholastic Aptitude Test, SAT)؛ آزمون وکالت چند ایالتی (Multistate Bar Examination, MBE).

طول آزمون لنگر

یکی از عوامل کلیدی در دستیابی به نتایج دقیق همترازسازی، طول آزمون لنگر است (مارنگو و همکاران[1]، ۲۰۱۸) که به هدف سنجش، ناهمگنی محتوای اندازه‌گیری شده و ویژگی‌های آزمون وابسته است (کولن و برنان، ۲۰۱۴). بسیاری از متخصصان با موضوع طول آزمون لنگر به‌منظور انتخاب حداکثر طول آزمون برای دستیابی به هدف‌های آماری و حداقل طول آزمون برای ملاحظاتی مانند امنیت آزمون مواجه شده‌اند (ریکر[2] و ون‌داویر، ۲۰۰۷). از دیدگاه یانگ و هوانگ (۱۹۹۶) یکی از عوامل تعیین‌کننده اهمیت همبستگی آزمون لنگر و آزمون کل، طول آزمون لنگر است. به‌طوری‌که با افزایش طول آزمون، همبستگی این دو آزمون افزایش می‌یابد و موجب بهبود همترازسازی می‌شود. هم‌چنین، بیدسکو (۱۹۸۵) اظهار نمود که طول نسبی آزمون لنگر، عاملی است که بر همبستگی آزمون لنگر و آزمون کل اثر می‌گذارد. سان‌ناسی[3] (۲۰۱۱، ص. ۹۰) در پژوهش خود بیان نمود که استفاده از آزمونی با سؤال‌های بیشتر، دقت برآوردهای همترازسازی را بهبود می‌بخشد، زیرا پایایی و همبستگی بین دو آزمون افزایش می‌یابد. کاهش تعداد سؤال‌ها در یک آزمون لنگر، همبستگی این دو آزمون را کاهش می‌دهد و درنتیجه، خطای همترازسازی افزایش می‌یابد (ریکر و ون‌داویر، ۲۰۰۷؛ یی[4]، ۲۰۰۹؛ پوهان، ۲۰۱۰؛ ژانگ[5] و کولن، ۲۰۱۳، لین و همکاران[6]، ۲۰۱۶).

نوع آزمون لنگر (ویژگی‌های آماری)

کولن و برنان (۲۰۰۴) معتقدند که آزمون لنگر باید به‌گونه‌ای طراحی شود که بتواند به‌دقت تفاوت‌های دو گروه را منعکس کند. انتظار می‌رود این آزمون که اغلب از آن به‌عنوان minitest یاد می‌شود، از نظر ویژگی‌های آماری و محتوایی مشابه آزمون کل باشد (سینهارای، ۲۰۱۷). سینهارای

1. Marengo et al.
2. Ricker
3. Sunnassee
4. Yi
5. Zhang
6. Lin et al.

و هالند (۲۰۰۶b) بیان نمودند که وقتی میانگین و پراکندگی دشواری سؤال‌های آزمون لنگر تقریباً برابر با آزمون کل باشد، می‌توان گفت این آزمون نماینده آماری آزمون کل است. از طرفی، در طراحی آزمون لنگر mini لازم است، سؤال‌های بسیار آسان و بسیار دشوار بکار رود تا از کفایت پراکندگی دشواری سؤال‌های این آزمون اطمینان حاصل شود. این انتخاب می‌تواند چالش‌هایی برای طراحان آزمون به دلیل فراوانی پایین‌تر آن‌ها ایجاد کند؛ بنابراین، برای پرداختن به این موضوع و به دلیل اهمیت انتخاب آزمون لنگر در طرح همترازسازی NEAT، سینهارای و هالند (۲۰۰۶a) یک آزمون لنگر را پیشنهاد نمودند که نماینده محتوای آزمون کل است و میانگین دشواری سؤال‌های آن با آزمون کل مشابه است؛ اما به دلیل استفاده از سؤال‌هایی با دشواری متوسط، پراکندگی دشواری سؤال‌های آن کمتر از آزمون کل است. سینهارای و هالند (۲۰۰۶a، ۲۰۰۶b، ۲۰۰۷) این نوع آزمون لنگر را miditest نامیدند. حال اگر مقدار پراکندگی دشواری سؤال‌های لنگر بین minitest و miditest باشد (کمتر از minitest و بیشتر از miditest)، از آن به‌عنوان semi-miditest نام بردند. نتایج پژوهش‌های سینهارای و هالند (۲۰۰۶a، ۲۰۰۷)، یی (۲۰۰۹)، لیو و همکاران (۲۰۱۱a، ۲۰۱۱b) و سینهارای (۲۰۱۷) نشان می‌دهد که همبستگی آزمون لنگر midi با آزمون کل از همبستگی آزمون لنگر mini با آزمون کل بیشتر است و مقدار همبستگی آزمون لنگر semi-midi بین این دو نوع آزمون قرار دارد؛ به‌عبارت‌دیگر، عملکرد آزمون لنگر midi نسبت به دیگر آزمون‌های لنگر بهتر است. این لنگر، دارای بالاترین مقدار همبستگی با آزمون کل است و نتایج همترازسازی آن دقیق‌تر گزارش‌شده است. پژوهش تری‌ویلر و همکاران (۲۰۱۶) نشان داد که آزمون لنگر midi همیشه بالاترین مقدار همبستگی را بین این سه نوع آزمون لنگر ندارد. با توجه به وجود این مطالعه که یافته‌های آن برخلاف پژوهش‌های ذکرشده است، سینهارای (۲۰۱۷) تصمیم گرفت این تفاوت را در پژوهشی بررسی نماید. او به دنبال پاسخ به این سؤال بود که آیا آزمون لنگر midi باید استفاده شود و در عمل ادامه یابد؟ یافته‌های این مطالعه نشان داد که برخلاف دیدگاه تری‌ویلر و همکارانش، لنگر midi همبستگی بین آزمون لنگر و آزمون کل را افزایش می‌دهد و عملکرد آن بهتر از آزمون mini است و باید از آن در همترازسازی آزمون‌ها استفاده کرد.

پایایی و آزمون لنگر

ازآنجایی‌که همترازسازی شکل قوی‌تر پیوند است (لیو و واکر[1]، ۲۰۰۷)، شرایطی برای همترازسازی باید در نظر گرفته شود تا بتوان پیوند بین دو آزمون را همترازسازی نامید. یکی از این شرایط، برابری پایایی است (لرد، ۱۹۸۰؛ انگاف، ۱۹۸۴؛ شی و نورسینی، ۱۹۹۵؛ دورانز و همکاران، ۲۰۱۰). این شرط برای همترازسازی دو فرم آزمون، ضروری و مهم تلقی می‌شود. یانگ و هوانگ (۱۹۹۶) بیان نمودند که همبستگی قوی بین آزمون لنگر و آزمون کل، نشانه‌ای از پایا و روا بودن آزمون لنگر است. ماسز و کیم (۲۰۰۷) در پژوهشی تأثیر نابرابری پایایی بر روش‌های همترازسازی با طرح NEAT را بررسی کردند. نتایج نشان داد که تفاوت در پایایی بین دو فرم آزمون، منجر به بیش برآورد تابع همترازسازی می‌شود. اگر توانایی افراد در گروه‌ها متفاوت و آزمون‌ها ناپایا باشند، نتایج همترازسازی در هر روشی با خطا همراه خواهد بود. همچنین، نتایج بیانگر آن بود که در روش تاکر، همبستگی آزمون لنگر و آزمون کل، از نظر پایایی آزمون لنگر قابل‌بررسی است. از دیدگاه بیدسکو (۱۹۸۵)، ژانگ و کولن (۲۰۱۳) و تری‌ویلر و همکاران (۲۰۱۶) یکی از عواملی که بر همبستگی این دو آزمون تأثیر می‌گذارد، پایایی آزمون کل است و بین این همبستگی و پایایی یک اثر متقابل وجود دارد. سینهارای و هالند (۲۰۰۶a) در مطالعه خود نشان دادند که پایایی آزمون لنگر midi بالاتر از لنگر mini است که عاملی برای بالاتر بودن همبستگی آزمون لنگر midi نسبت به لنگر mini است. از طرفی، نتایج پژوهش ریکر و ون‌داویر (۲۰۰۷) نشان می‌دهد که کاهش تعداد سؤال‌های لنگر باعث کاهش پایایی آن و درنتیجه کاهش همبستگی آزمون لنگر با آزمون کل می‌شود؛ بنابراین، از دیدگاه آن‌ها، یکی از عوامل کنترل خطای همترازسازی دو فرم آزمون، پایایی آزمون لنگر است.

محتوای آزمون لنگر

آزمون لنگر باید تا حد امکان ویژگی‌های محتوایی مشابه آزمون کل داشته باشد و سؤال‌های آن معرف محتوای آزمون باشد، چراکه به گفته کلاین و جارجورا[2] (۱۹۸۵) عدم تناسب محتوایی در آزمون لنگر، همترازسازی را تحت تأثیر قرار می‌دهد و منجر به ایجاد خطای همترازسازی می‌شود. از

[1]. Walker
[2]. Klein & Jarjoura

دیدگاه یانگ و هوانگ (۱۹۹۶) و تری‌ویلر و همکاران (۲۰۱۶) یک عامل مهم در همبستگی آزمون لنگر و آزمون کل این است که آزمون لنگر نماینده‌ای از محتوای آزمون کل باشد. اگر دو فرم آزمون، سازهٔ مشابهی را اندازه بگیرند و محتوای مشابهی داشته باشند، همبستگی بین آزمون لنگر و آزمون کل افزایش می‌یابد (لین و همکاران، ۲۰۱۶؛ مارنگو و همکاران، ۲۰۱۸)؛ بنابراین، یکی از راه‌های دستیابی به همبستگی بالاتر بین این دو آزمون، دقت در انتخاب محتوای آزمون لنگر و تشابه محتوای آن با آزمون کل است (بیدسکو، ۱۹۸۵؛ سینهارای، ۲۰۱۷). بالا[1] (۱۹۸۸)، سینهارای و هالند (۲۰۰۶a) و لین و همکاران (۲۰۱۶) معتقدند که اگر آزمون لنگر نماینده محتوای آزمون کل نباشد، توانایی متفاوتی اندازه‌گیری می‌شود. درنتیجه همبستگی بین دو آزمون کاهش می‌یابد و منجر به سوگیری در نتایج همترازسازی می‌شود.

توانایی گروه‌های آزمودنی

هنگام استفاده از طرح NEAT برای فرایند همترازسازی، اگر دو گروه از آزمودنی‌ها از نظر توانایی بسیار متفاوت باشند، نتایج همترازسازی با سوگیری همراه خواهد شد (لیو و همکاران، ۲۰۰۹؛ ماسز و همکاران، ۲۰۱۰؛ ون‌درلیندن[2] و ویبرگ، ۲۰۱۰؛ هگ[3]، ۲۰۱۰؛ لیو و همکاران، ۲۰۱۱a، ۲۰۱۱b؛ آریکان و جلبال[4]، ۲۰۱۸). سو و همکاران[5] (۲۰۰۹) گزارش نمودند که اگر بر تفاوت توانایی دو گروه از آزمودنی‌ها افزوده شود، همبستگی بین آزمون لنگر با آزمون کل کاهش می‌یابد. تحت این شرایط نتایج روش همترازسازی poststratification نسبت به روش همترازسازی زنجیره‌ای دارای سوگیری بیشتری است (پوهان، ۲۰۱۰). از دیدگاه یی (۲۰۰۹)، اگر تفاوت توانایی آزمودنی‌ها در دو گروه زیاد باشد، همبستگی بین آزمون لنگر و آزمون کل برای سه نوع لنگر (mini، midi، semi-midi) تفاوت چندانی ندارد. سینهارای (۲۰۱۷) در مطالعه خود دریافت که اگر تفاوت توانایی آزمودنی‌ها کم باشد، نتایج تمام روش‌های همترازسازی رضایت‌بخش است و شرایط لنگر خیلی مهم نیست؛ اما اگر تفاوت در توانایی افراد زیاد باشد، ویژگی‌های مختلف آزمون لنگر اهمیت ویژه‌ای پیدا می‌کند.

روش‌های همترازسازی

با توجه به مدل‌های روان‌سنجی، رویکردهای همترازسازی مختلفی توسعه‌یافته است که به‌طورکلی می‌توان آن‌ها را به دو طبقه رویکردهای همترازسازی مبتنی بر نظریه کلاسیک آزمون و نظریه سؤال ـ پاسخ تقسیم کرد. هنگام استفاده از طرح NEAT برای همترازسازی بر اساس CTT، رویکردهای همترازسازی شامل همترازسازی خطی (همترازسازی خطی زنجیره‌ای[6]، همترازسازی تاکر[7]، همترازسازی لوین[8]) و همترازسازی همصدک (همترازسازی همصدک زنجیره‌ای، همترازسازی بران ـ هالند[9] و برآورد فراوانی[10]) است (برنان و همکاران، ۲۰۰۹؛ کولن و برنان، ۲۰۱۴؛ گنزالس و ویبرگ، ۲۰۱۷). ون‌داویر (۲۰۰۸) در پژوهشی نشان داد که هرگاه میانگین و انحراف استاندارد هر دو نمونه در آزمون لنگر برابر باشد، نتایج همه روش‌های همترازسازی یکسان است. از طرفی، اگر همبستگی آزمون لنگر و آزمون کل خیلی پایین باشد، نمره‌های هم‌تراز شده قابل‌معاوضه نیستند. به همین ترتیب، اگر همبستگی آزمون لنگر با آزمون کل بالا باشد، نتایج روش‌های همترازسازی زنجیره‌ای، لوین و تاکر مشابه است. در پژوهش ماسز و کیم (۲۰۰۷) اشاره‌شده است که در روش همترازسازی تاکر با افزایش همبستگی آزمون لنگر با آزمون کل، میزان خطای استاندارد همترازسازی کاهش می‌یابد. از دیدگاه ریکر و ون‌داویر (۲۰۰۷) نتایج روش‌های همترازسازی کرنل و برآورد فراوانی به همبستگی بین آزمون لنگر و آزمون کل بستگی دارد. بالا (۱۹۸۸) بیان نمود که کاهش همبستگی آزمون لنگر و آزمون کل منجر به کاهش عملکرد همترازسازی می‌شود. علاوه بر این، با کاهش مقدار این همبستگی، تفاوت بین نتایج روش‌های همترازسازی همصدک و خطی افزایش پیدا می‌کند. پوهان (۲۰۱۰) در پژوهشی روش‌های همترازسازی تاکر، لوین و زنجیره‌ای را تحت شرایط مختلف مقایسه کرد. نتایج نشان داد که اگر همبستگی

1. Balla
2. van der Linden & Wiberg
3. Hagge
4. Arikan & Gelbal
5. Suh et al.
6. chained
7. Tucker
8. Levine
9. Braun & Holland
10. frequency estimation

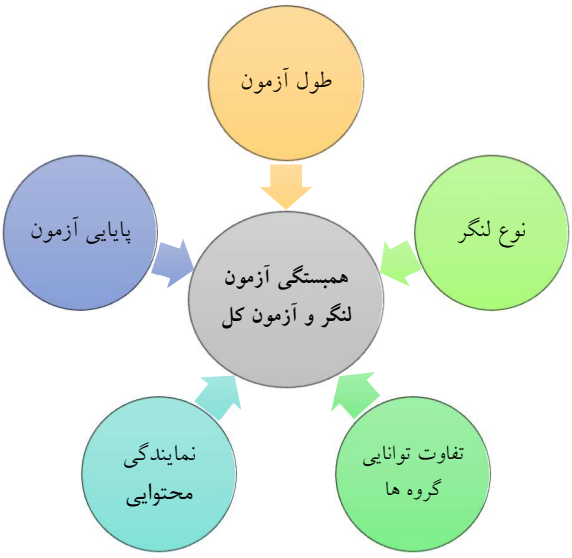
بین آزمون لنگر و آزمون کل بسیار کم باشد، هیچ‌یک از روش‌های همترازسازی نتیجه مناسبی ارائه نمی‌دهد. این سه روش زمانی قابل‌استفاده هستند که تفاوت در توانایی آزمودنی‌ها کم و مقدار همبستگی آزمون لنگر و آزمون کل نسبتاً بالا (۰/۷ به بالا) باشد. اگر گروه‌ها متفاوت و مقدار همبستگی آزمون لنگر و آزمون کل بالای ۰/۹ باشد، روش همترازسازی تاکر مناسب است. نتایج پژوهش سو و همکاران (۲۰۰۹) در ارزیابی عملکرد روش‌های مختلف همترازسازی نشان داد که افزایش تفاوت بین توانایی آزمودنی‌ها در دو فرم به کاهش همبستگی آزمون لنگر و آزمون کل منجر می‌شود. به گفته آن‌ها تأثیر تفاوت توانایی آزمودنی‌ها بر نتایج همترازسازی بیشتر از تفاوت بین فرم‌های آزمون است. اگر تشابه بین گروه‌ها زیاد و بین آزمون‌ها کم باشد، روش همترازسازی تاکر مناسب‌ترین روش است و در صورت معکوس شدن شرایط، روش همترازسازی لوین مناسب‌تر است.

بحث

این مقاله، اولین مطالعه در حوزه سنجش به‌ویژه موضوع همترازسازی است که از روش‌های مرور سیستماتیک برای پرداختن به سؤال‌های پژوهش جهت بررسی اثر همبستگی آزمون لنگر و آزمون کل بر نتایج همترازسازی تحت طرح گروه‌های نامعادل با آزمون لنگر استفاده نمود. راهبردهای جستجوی جامعی برای یافتن مطالعه‌هایی که ملاک‌های در نظر گرفته‌شده را داشته باشند، بکار گرفته شد. به کمک این مرور، ۱۸ مطالعه منتشرشده از اجرای همترازسازی تحت طرح NEAT به‌تفصیل تحلیل شد تا تأثیر همبستگی بر یافته‌های همترازسازی تعیین شود.

سؤال ۱: کدام عوامل بر همبستگی بین آزمون لنگر و آزمون کل مؤثر هستند؟

هوانگ و یانگ (۱۹۹۶) بیان کردند که دو عاملی که اهمیت همبستگی آزمون لنگر با آزمون کل را تعیین می‌کند، نمایندگی محتوایی آزمون لنگر و طول آن است. همبستگی بالا بین این دو آزمون نشانهٔ پایایی و روایی آزمون لنگر است و طول بیشتر آن منجر به افزایش این همبستگی و درنتیجه بهبود همترازسازی می‌شود. طبق نظر بیدسکو (۱۹۸۵)، پایایی آزمون کل و طول نسبی هر دو آزمون، دو عامل مؤثر بر همبستگی آزمون لنگر و آزمون کل است. علاوه بر این، دقت در انتخاب محتوای آزمون لنگر باعث می‌شود تا همبستگی بالاتری به دست آید. با توجه به اهمیت همبستگی این دو آزمون، هابرمن و دورانز (۲۰۰۹) گزارش نمودند که سه عامل بر رانش مقیاس تحت طرح NEAT تأثیر می‌گذارد. این سه عامل، تفاوت زیاد در توانایی گروه‌ها، محتوای متفاوت آزمون لنگر و آزمون کل و فقدان همبستگی بالا بین این دو آزمون است. ژانگ و کولن (۲۰۱۳) معتقدند که دقت همترازسازی مستقیماً با همبستگی بین آزمون لنگر و آزمون کل مرتبط است و این همبستگی از پایایی آزمون کل، طول آزمون لنگر و طول آزمون کل تأثیر می‌پذیرد و این عوامل، خطای همترازسازی را تحت تأثیر قرار می‌دهند. تری‌ویلر و همکاران (۲۰۱۶) بیان کردند که همبستگی آزمون لنگر با آزمون کل، نه‌تنها به پراکندگی دشواری سؤال‌های آزمون، بلکه به عوامل مستقلی مانند پایایی آزمون لنگر و همبستگی بین نمره واقعی آزمون لنگر و آزمون کل نیز وابسته است. مستقل بودن این دو عامل به این معناست که اگر چند سؤال با ویژگی‌های مشابه به آزمون اضافه شود، پایایی آزمون بهبود می‌یابد، ولی همبستگی تغییری نمی‌کند. علاوه بر این عوامل، نوع لنگر نیز بر این رابطه اثر می‌گذارد. به‌طوری‌که آزمون لنگر midi نسبت به آزمون لنگر mini همبستگی بالاتری با آزمون کل دارد (سینهارای و هالند، ۲۰۰۶a، ۲۰۰۷؛ یی، ۲۰۰۹؛ لیو و همکاران، ۲۰۱۱a، ۲۰۱۱b؛ سینهارای، ۲۰۱۷). عامل دیگری که بر همبستگی این دو آزمون تأثیر می‌گذارد، ساختار محتوایی آزمون لنگر و تشابه آن با آزمون کل است (بیدسکو، ۱۹۸۵؛ یانگ و هوانگ، ۱۹۹۶؛ سینهارای و هالند، ۲۰۰۶a؛ تری‌ویلر و همکاران، ۲۰۱۶؛ سینهارای، ۲۰۱۷؛ مارنگو و همکاران، ۲۰۱۸). تفاوت محتوایی بین آزمون لنگر و آزمون کل به کاهش همبستگی و بروز سوگیری در نتایج همترازسازی منجر می‌شود (بالا، ۱۹۸۸؛ لین و همکاران، ۲۰۱۶). از سوی دیگر، زمانی که سازه‌های دو فرم آزمون از نظر محتوا مشابه، همبستگی بین دو فرم بالا و تابع پیوند آن‌ها تغییرناپذیر باشد، شرایطی ایجاد می‌شود که می‌توان نمره‌های هم‌تراز شده را معاوضه نمود (دورانز، ۲۰۰۴؛ به نقل از لین و همکاران، ۲۰۱۶، ص. ۲). هرگاه تفاوت در توانایی دو گروه زیاد باشد، همبستگی آزمون لنگر و آزمون کل کاهش می‌یابد (سو و همکاران، ۲۰۰۹؛ یی، ۲۰۰۹؛ پوهان، ۲۰۱۰؛ سینهارای، ۲۰۱۷). با توجه به آنچه ذکر شد، می‌توان گفت ۵ عامل بر همبستگی بین آزمون لنگر و آزمون کل مؤثر است. این عوامل که در شکل ۲ نشان داده‌شده است عبارت‌اند از: طول آزمون (لنگر، کل)، پایایی آزمون (لنگر، کل)، نوع لنگر از نظر ویژگی‌های آماری، ساختار محتوایی آزمون لنگر (نمایندگی محتوایی) و تفاوت در توانایی هر دو گروه.



شکل ۲. عوامل مؤثر بر همبستگی آزمون لنگر و آزمون کل

سؤال ۲: همبستگی آزمون لنگر و آزمون کل چه اثری بر نتایج همترازسازی دارد؟

یکی از مهم‌ترین عوامل مؤثر بر کارایی همترازسازی دو فرم آزمون، همبستگی آزمون لنگر با آزمون کل است (بیدسکو، ۱۹۸۵؛ ون‌داویر، ۲۰۰۸؛ سینهارای و هالند، ۲۰۰۶a). از نظر لرد (۱۹۷۵) این همبستگی در حذف سوگیری مؤثر است (به نقل از بالا، ۱۹۸۸، ص. ۴۱۰). یانگ و هوانگ (۱۹۹۶) بیان نمودند اگر مقدار همبستگی این دو آزمون ۰/۹۹۹ باشد، فرایند همترازسازی دارای نتایج یکسانی است. البته به گفته سینهارای و هالند (۲۰۰۶a) این مقدار همبستگی در عمل اتفاق نمی‌افتد. از دیدگاه سان‌ناسی (۲۰۱۱)، وجود همبستگی ۰/۸ یا بیشتر، بین آزمون لنگر و آزمون کل، عامل مهمی در موفقیت فرایند همترازسازی است. کاهش این همبستگی باعث کاهش عملکرد روش‌های همترازسازی و عدم معاوضه نمره‌ها می‌شود (بالا، ۱۹۸۸؛ ون‌داویر، ۲۰۰۸؛ لین و همکاران، ۲۰۱۶). نتایج روش‌های همترازسازی خطی، همصدک، زنجیره‌ای، تاکر، لوین و کرنل تحت طرح گروه‌های نامعادل با آزمون لنگر، تحت تأثیر مقدار این همبستگی قرار دارد (بالا، ۱۹۸۸؛ ریکر و ون‌داویر، ۲۰۰۷؛ ماسز و کیم، ۲۰۰۷؛ ون‌داویر، ۲۰۰۸، پوهان، ۲۰۱۰). اگر همبستگی آزمون لنگر و آزمون کل بالا باشد، نتایج همترازسازی روش‌های لوین، تاکر و زنجیره‌ای مشابه است (ون‌داویر، ۲۰۰۸؛ پوهان، ۲۰۱۰). پوهان (۲۰۱۰) هم‌چنین خاطرنشان کرد که اگر تفاوت بین گروه‌ها کم باشد، برای به دست آوردن یک فرایند همترازسازی دقیق در این سه روش، به همبستگی بالاتر از ۰/۷ نیاز است. اگر توانایی گروه‌ها متفاوت باشد، روش زنجیره‌ای مقدار خطای همترازسازی کمتری ایجاد می‌کند؛ اما اگر توانایی گروه‌ها با یکدیگر بسیار متفاوت و همبستگی آزمون لنگر و آزمون کل ۰/۹ یا بیشتر باشد، در بین این سه روش، روش تاکر مناسب‌ترین روش همترازسازی است. علاوه بر نوع روش همترازسازی، نوع آزمون لنگر از نظر ویژگی‌های آماری بر صحت نتایج همترازسازی مؤثر است. یافته‌های برخی از مطالعه‌ها نشان می‌دهد که همبستگی آزمون لنگر midi با آزمون کل بیشتر از آزمون لنگر mini است و این عامل باعث می‌شود که نتایج همترازسازی با آزمون لنگر midi معمولاً دقیق‌تر از آزمون لنگر mini باشد (سینهارای و هالند، ۲۰۰۶a؛ لیو و همکاران، ۲۰۰۹؛ یی، ۲۰۰۹؛ لیو و همکاران، ۲۰۱۱a، ۲۰۱۱b؛ سینهارای و همکاران، ۲۰۱۲؛ سینهارای، ۲۰۱۷). بالا (۱۹۸۸) معتقد است که کاهش همبستگی آزمون لنگر و آزمون کل باعث می‌شود تا عملکرد روش‌های همترازسازی نیز کاهش یابد. در این شرایط، افرادی که به آزمون دشوارتر پاسخ داده‌اند، متضرر می‌شوند. از دیدگاه ریکر و ون‌داویر (۲۰۰۷) همبستگی این دو آزمون و پایایی آزمون لنگر، دوعاملی هستند که خطای همترازسازی را مدیریت می‌کند. افزایش همبستگی آزمون لنگر و آزمون کل باعث کاهش خطای همترازسازی و افزایش دقت در نتایج می‌شود، زیرا بین همبستگی این دو آزمون و خطای استاندارد همترازسازی رابطه معکوس وجود دارد (سینهارای و هالند، ۲۰۰۶a، ۲۰۰۷؛ ماسز و کیم، ۲۰۰۷؛ ژانگ و کولن، ۲۰۱۳؛ تری‌ویلر و همکاران، ۲۰۱۶؛ مارنگو و همکاران، ۲۰۱۸). بر اساس این مطالب، دقت نتایج همترازسازی

تحت طرح NEAT مستقیماً به همبستگی آزمون لنگر و آزمون کل مرتبط است. به‌طوری‌که افزایش این همبستگی، منجر به کاهش خطای استاندارد همترازسازی و درنتیجه افزایش دقت در نمره‌های هم‌تراز شده می‌شود؛ بنابراین، یکی از مؤلفه‌های کلیدی و مهم برای اجرای یک همترازسازی موفق تحت این طرح، وجود همبستگی بالا بین آزمون لنگر و آزمون کل است. از طرفی، با توجه به مطالعه‌های بررسی‌شده در این مرور، مقدار این همبستگی بر روش‌های مختلف همترازسازی و نتایج آن‌ها مؤثر است.

با توجه به ملاک‌های مورداستفاده برای ورود مطالعه‌ها، دو محدودیت برای این مرور ایجاد شد. یک محدودیت این است که فقط مطالعه‌هایی که از دو فرم آزمون برای فرایند همترازسازی استفاده کرده بودند، در نظر گرفته شد؛ به‌عبارت‌دیگر، مطالعه‌هایی که شامل آزمون‌های چندگانه، آزمونچه و آزمون‌های چندوجهی بودند، بررسی نشدند. از سوی دیگر، سؤال‌های مورد تحلیل در مطالعه‌ها از نوع دو ارزشی بود و داده‌های چند ارزشی در این مرور لحاظ نشد؛ بنابراین، یافته‌های این مطالعه به پژوهش‌هایی محدود شد که شرایط مختلف فرایند اجرای همترازسازی برای دو فرم آزمون را تحت طرح NEAT و داده‌های دو ارزشی بررسی نمودند. در پژوهش‌های آتی، آزمون‌های چندگانه و آزمون‌هایی با داده‌های چند ارزشی را می‌توان برای بررسی اثر همبستگی بین آزمون لنگر و آزمون کل در نظر گرفت. طبق نظر ریکر و ون‌داویر (۲۰۰۷) یکی از چالش‌های طراحان آزمون، آزمون لنگر طولانی برای دست یافتن به ویژگی‌های آماری و آزمون لنگر کوتاه برای حفظ ملاحظاتی مانند امنیت و زمان آزمون است. با توجه به یافته‌های این مرور، یکی از عوامل مؤثر بر همبستگی آزمون لنگر و آزمون کل، طول آزمون لنگر است. برخی از مطالعه‌ها این دو مؤلفه را باهم بررسی کردند که نتایج نشان داد با افزایش طول آزمون لنگر، مقدار همبستگی نیز افزایش می‌یابد. بر این اساس، در مطالعه‌های آینده لازم است طول بهینه آزمون لنگر برای حفظ تناسب محتوا و همبستگی آن با آزمون کل بررسی شود. همان‌طور که در این مرور نشان داده شد، این مطالعه‌ها بر اساس ملاک‌های ورود دارای ویژگی‌های متفاوتی بودند. همین امر باعث گردید تا در پاسخ به سؤال‌ها، شرایط متفاوتی در نظر گرفته شود. این موضوع یکی از نقاط قوت این مرور محسوب می‌شود، چراکه به شناسایی عوامل مربوط به همبستگی آزمون لنگر و آزمون کل و تأثیر آن بر همترازسازی از جنبه‌های مختلف کمک نمود.

نتیجه‌گیری

در این مرور، اثر همبستگی آزمون لنگر و آزمون کل بر نتایج همترازسازی تحت طرح گروه‌های نامعادل با آزمون لنگر بررسی گردید و عواملی که بر این همبستگی تأثیر می‌گذارند، شناسایی شد. برای دستیابی به این هدف‌ها، ۱۸ مطالعه در زمینهٔ همترازسازی با طرح NEAT مرور شد. مطالعه‌های واردشده به این مرور بر اساس جایگاه و موقعیت سؤال‌های لنگر، ویژگی‌های آماری آزمون لنگر، نوع داده‌ها (واقعی، شبیه‌سازی)، نوع توزیع داده‌ها (نرمال، دارای کجی منفی)، روش تحلیل داده‌ها (CTT، IRT، Rasch و کرنل) و اندازه نمونه (از ۲۵ تا ۲۰۰۰۰ آزمودنی) دارای شرایط متفاوتی بودند. بررسی‌ها حاکی از آن بود که باوجود شرایط متفاوت بین مطالعه‌های این مرور، از نظر اهمیت همبستگی آزمون لنگر و آزمون کل بر بهبود فرایند همترازسازی و عوامل مؤثر بر این همبستگی، نتایج مشابهی برقرار است. از میان این مطالعه‌ها، پژوهش بالا (۱۹۸۸) و سینهارای و هالند (۲۰۰۶a) به‌طور خاص به بررسی همبستگی بین آزمون لنگر و آزمون کل و تأثیر آن بر فرایند همترازسازی پرداخته‌اند. در سایر مطالعه‌ها، موضوع همبستگی آزمون لنگر و آزمون کل در کنار مؤلفه‌های دیگر (طول لنگر، نوع لنگر، پایایی آزمون‌ها، مقایسه روش‌ها) بررسی‌شده است. در این مطالعه‌ها مشاهده شد که همبستگی بالا بین آزمون لنگر و آزمون کل، یکی از شروط لازم برای اجرای همترازسازی کارآمد است. در مطالعه‌های بررسی‌شده، مقدار این همبستگی اغلب در دامنه ۰/۷ تا ۰/۹ در نظر گرفته و برآورد شده است؛ بنابراین، برای آن‌که فرایند همترازسازی در برآورد پارامترها و نمره‌ها از دقت کافی برخوردار باشد، لازم است مقدار همبستگی آزمون لنگر و آزمون کل ۰/۷ یا بیشتر باشد. مقدار این همبستگی، شاخصی از صحت نتایج همترازسازی است که با افزایش مقدار آن، کیفیت و دقت برآورد پارامترها در گروه‌ها افزایش و خطای استاندارد همترازسازی کاهش می‌یابد. علاوه بر این، نتایج بیانگر آن است که مقدار همبستگی بر روش‌های همترازسازی تأثیر دارد. هنگامی‌که آزمون لنگر و آزمون کل، همبستگی بسیار کمی داشته باشند، هیچ‌یک از روش‌های همترازسازی نتایج خوبی ایجاد نمی‌کند و نمره‌های هم‌تراز شده قابل‌معاوضه نیستند. اجرای همترازسازی در این شرایط توصیه نمی‌شود، زیرا عملکرد روش‌های همترازسازی با کاهش همبستگی کاهش می‌یابد و نتایج به نفع آزمودنی‌های شرکت‌کننده در آزمون آسان‌تر خواهد بود. همچنین، با توجه به اهمیت همبستگی بین آزمون لنگر و آزمون کل بر فرایند همترازسازی و وجود ارتباط مستقیم بین مقدار این همبستگی و دقت در همترازسازی، در این مرور، عوامل مؤثر بر این

همبستگی، موردبررسی قرار گرفت. یکی از عواملی که در صحت فرایند همترازسازی نقش بسزایی دارد، طول آزمون لنگر است. یافته‌های این مرور نشان داد که با افزایش طول آزمون لنگر، مقدار همبستگی این دو آزمون افزایش می‌یابد و به دنبال آن نتایج همترازسازی بهبود می‌یابد. یکی از شروط لازم برای اجرای فرایند همترازسازی، برابری پایایی دو آزمون است. نتایج مرور حاکی از آن بود که بین پایایی آزمون (لنگر و کل) و همبستگی این دو آزمون تقابل وجود دارد. به‌گونه‌ای که با افزایش پایایی، همبستگی آزمون لنگر و آزمون کل افزایش می‌یابد. با توجه به ویژگی‌های آماری آزمون لنگر، شواهد گویای آن بود که نوع لنگر (semi-midi, midi, mini) بر همبستگی این دو آزمون و دستیابی به نتایج دقیق‌تر مؤثر است. یک‌راه برای دستیابی به همبستگی بالا بین آزمون لنگر و آزمون کل، وجود تشابه محتوایی بین این دو آزمون است؛ بنابراین، با دقت در انتخاب سؤال‌های آزمون لنگر از نظر محتوا می‌توان به نتایج دقیق‌تری در فرایند همترازسازی دست‌یافت. در کنار این عوامل، هنگامی‌که توانایی آزمودنی‌ها در دو گروه متفاوت باشد، یافته‌های حاصل از فرایند همترازسازی با سوگیری همراه خواهد شد، زیرا این تفاوت بر همبستگی آزمون لنگر با آزمون کل اثر گذاشته و منجر به کاهش مقدار آن می‌شود. بر این اساس، لازم است پیش از اجرای فرایند همترازسازی، توانایی گروه‌های آزمودنی ارزیابی شود تا فرایند همترازسازی با حداقل سوگیری اجرا گردد. بنا بر آنچه مطرح شد، می‌توان گفت که تعداد سؤال‌های آزمون (لنگر، کل)، پایایی آزمون (لنگر، کل)، پراکندگی دشواری سؤال‌های آزمون لنگر، تشابه محتوای آزمون لنگر با آزمون کل و تفاوت در توانایی آزمودنی‌های دو گروه، مجموعه عواملی هستند که بر همبستگی آزمون لنگر با آزمون کل تأثیر می‌گذارند. شناسایی این عوامل، راه را برای طراحان آزمون، تحلیل‌گران آزمون و افرادی که قصد اجرای فرایند همترازسازی دارند، تسهیل می‌نماید. از سوی دیگر، بستر بسیار مناسبی را برای سیاست‌گذاران و مدیران سازمان سنجش آموزش کشور برای طراحی انواع آزمون‌ها و اجرای همترازسازی فراهم می‌آورد. با توجه به نتایج مرور، طراحان با شناخت و درک عوامل مؤثر بر همبستگی آزمون لنگر و آزمون کل می‌توانند هنگام طراحی فرم‌های مختلف از یک آزمون، آن‌ها را مدنظر قرار دهند تا در فرایند همترازسازی به نتایج بهتری دست یابند و با حداقل خطا مواجه شوند. یافته‌های مرور بیانگر آن است که لنگر midi از لنگر semi-midi و mini با آزمون کل همبستگی بالاتری دارد و عملکرد آن در بهبود نتایج همترازسازی بهتر از دو نوع لنگر دیگر است. این شواهد به طراحان آزمون کمک می‌کند تا آزمون لنگر midi و یا semi-midi را انتخاب کنند و با اطمینان از کیفیت و دقت کافی در برآورد پارامترها و نمره‌های آزمون‌های هم‌تراز شده، در زمان و هزینه ساخت آزمون لنگر نیز صرفه‌جویی نمایند. بااین‌همه، انتخاب نوع لنگر به هدف‌ها و شرایط اجرای آزمون بستگی دارد. علاوه بر شناسایی عوامل مؤثر بر همبستگی بین آزمون لنگر و آزمون کل که در مرور به آن پرداخته شد، اثر این همبستگی بر نتایج همترازسازی نیز موردبررسی قرار گرفت. یافته‌ها بیانگر آن بود که نتایج فرایند همترازسازی تحت طرح NEAT برای روش‌های خطی (لوین، تاکر، زنجیره‌ای)، همصدک (برآورد فراوانی، زنجیره‌ای) و کرنل به مقدار همبستگی این دو آزمون وابسته است. به‌طوری‌که در مقادیر بالای همبستگی، این روش‌ها از نتایج مشابهی برخوردار هستند. حال اگر مقدار این همبستگی با توجه عوامل ذکرشده، کم شود، عملکرد روش‌های همترازسازی نیز کاهش می‌یابد و در پی آن امکان معاوضه نمره‌ها فراهم نخواهد بود. از طرفی، بین همبستگی این دو آزمون و خطای همترازسازی رابطه معکوسی برقرار است. به‌گونه‌ای که با افزایش مقدار همبستگی آزمون لنگر با آزمون کل، خطای همترازسازی کاهش می‌یابد و نتایج حاصل از فرایند همترازسازی دقیق‌تر می‌شود. با توجه به اهمیت تأثیر این همبستگی بر فرایند همترازسازی که در مرور به آن اشاره شد، لازم است طراحان آزمون و مجریان برگزارکننده، مقدار این همبستگی و عوامل مؤثر بر آن را در مراحل ساخت آزمون و قبل از انجام تحلیل‌های مرتبط با همترازسازی به‌دقت بررسی و تحلیل نمایند تا از بروز خطای همترازسازی و سوگیری در نتایج کاسته شود.

References

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). American Council on Education.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Educational Testing Service.

Arikan, C. A., & Gelbal, S. (2018). The effect of mini and midi anchor tests on test equating. *The International Journal of Progressive Education, 14*(2), 148-160. https://doi.org/10.29329/ijpe.2018.139.11

Balla, J. (1988). The effects of reducing correlation of external anchors on test equating methods for the equivalent groups and non-equivalent groups designs. *International Journal of Educational Research, 12*(4), 409-425. https://doi.org/10.1016/0883-0355(88)90034-1

Braun, H. I., & Holland, P. W. (1982). Observed score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 9-49). Academic.

Brennan, R. L., Wang, T., Kim, S., & Seol, J. (2009). *Equating recipes*. CASMA. https://education.uiowa.edu/sites/education.uiowa.edu/files/2021-11/casma-monograph-1.pdf

Budescu, D. (1985). Efficiency of linear equating as a function of the length of the anchor test. *Educational Measurement, 22*(1), 13–20. https://www.jstor.org/stable/1434562

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, *28*(4), 227–246. https://doi.org/10.1177/0146621604265031

Dorans, N. J., Liu, J., & Hammond, S. (2008). Anchor test type and population invariance: An exploration across subpopulations and test administrations. *Applied Psychological Measurement, 32*(1), 81–97. https://doi.org/10.1177/0146621607311580

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). *Principles and practices of test score equating* (RR-10-29). ETS. https://files.eric.ed.gov/fulltext/ED523737.pdf

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2011). Equating test scores: toward best practices. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling and linking* (pp. 21-42). Springer.

Fenton, L., Lauckner, H., & Gilbert, R. (2015). The QATSDD critical appraisal tool: comments and critiques. *Evaluation in clinical Practice, 21*, 1125-1128. https://doi.org/10.1111/jep.12487

Gonzalez, J., & Wiberg, M. (2017). *Applying test equating method using R*. Springer.

Haberman, S., & Dorans, N. J. (2009, April). *Scale consistency, drift, stability: Definitions, distinctions, and principles* [Paper presentation]. National Council on Measurement in Education, San Diego, CA. http://www.ets.org/legal/index.html

Hagge, S. L. (2010). *The impact of equating method and format representation of common items on the adequacy of mixed-format test equating using nonequivalent groups* (Doctoral Dissertation, University of Iowa). https://doi.org/10.17077/etd.bc5ticit

Kanamori, L. F., Xu, C., Hasan, S. S., & Doi, S. A. (2021). Quality versus risk of bias assessment in clinical research. *Clinical Epidemiology, 129*, 172-175. https://doi.org/10.1016/j.jclinepi.2020.09.044

Klein, L. W., & Jarjoura, D. (1985). The importance of content representation for common-item equating with nonrandom groups. *Educational Measurement, 22*(3), 197–206. http://www.jstor.org/stable/1435033

Kolen, M. J. (2020). Equating with small samples (Commentary). *Applied Measurement in Education, 33*(1), 77-82. https://doi.org/10.1080/08957347.2019.1674308

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). Springer.

Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scaling, and linking* (3rd ed.). Springer.

Lasserson, T. J., Thomas, J., & Higgins, J. P. T. (2019). Starting a review. In J. P. T. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page & V. A. Welch (Eds.), *Cochrane Handbook for systematic review of interventions* (2nd ed., pp. 1-12). Wiley-Blackwell.

Lin, P., Dorans, N., & Weeks, J. (2016). *Linking composite scores: Effects of anchor test length and content representativeness* (Research Report No. RR-16-36). Educational Testing Service. https://doi.org/10.1002/ets2.12122

Liu, J., Sinharay, S., Holland, P. W., Feigenbaum, M., & Curley, E. (2009). *The effects of different types of anchor tests on observed score equating*. ETS. https://www.ets.org/research/contact.html

Liu, J., Sinharay, S., Holland, P. W., Feigenbaum, M., & Curley, E. (2011a). Test score equating using a mini-version anchor and a midi anchor: A case study using SAT data. *Educational Measurement, 48*(4), 361-379. https://doi.org/10.1111/j.1745-3984.2011.00150.x

Liu, J., Sinharay, S., Holland, P. W., Feigenbaum, M., & Curley, E. (2011b). Observed score equating using a mini-version anchor and an anchor with less spread of difficulty: A comparison study. *Educational and Psychological Measurement, 71*, 346–361. https://doi.org/10.1177/0013164410375571

Liu, J., & Walker, M. E. (2007). Score linking issues related to test content changes. In N. J. Dorans, M. Pommerich & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 109-134). Springer.

Livingston, S. A. (2004). *Equating test scores (without IRT)*. ETS. https://www.ets.org/Media/Research/pdf/LIVINGSTON.pdf

Lord, F. M. (1975). *A survey of equating methods based on item characteristic curve theory* (RB 75-13). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1975.tb01052.x

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Educational Measurement, 14*(2), 117-138. http://doi.org/10.2307/1434011

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Lawrence Erlbaum.

Marengo, D., Miceli, R., Rosato, R., & Settanni, M. (2018). Placing multiple tests on a common scale using a post-test anchor design: Effects of item position and order on the stability of parameter estimates. *Applied Mathematics and Statistics, 4,* 1-14. http://doi.org/10.3389/fams.2018.00050

Moses, T., Deng, W., & Zhang, Y. L. (2010). *The use of two anchors in nonequivalent groups with anchor test (NEAT) equating*. ETS. http://doi.org/10.1002/j.2333-8504.2010.tb02230.x

Moses, T., & Kim, S. (2007). *Reliability and the nonequivalent groups with anchor test design* (RR-07-16). ETS. https://doi.org/10.1002/j.2333-8504.2007.tb02058.x

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S.,... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Page, M. J., McKenzie, J. E., & Higgins, J. P. T. (2018). Tools for assessing risk of reporting biases in studies and syntheses of studies: A systematic review. *BMJ open, 8*(3), 1-16. https://doi.org/10.1136/bmjopen-2017-019703

Petersen, N. S. (2007). Equating: best practices and challenges to best practices. In N. J. Dorans, M. Pommerich & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 59-72). Springer.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). Macmillan.

Petersen, N. S., Marco, G. L., & Stewart, E. E. (1982). A test of the adequacy of linear score equating models. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 71–135). Academic.

Puhan, G. (2010). A comparison of chained linear and poststratification linear equating under different testing conditions. *Educational Measurement, 47*(1), 54-75. https://doi.org/10.1111/j.1745-3984.2009.00099.x

Ricker, K. L., & von Davier, A. A. (2007). *The Impact of anchor test length on equating results in a nonequivalent groups design*. ETS. https://www.ets.org/research/contact.html

Ryan, J., & Brockmann, F. (2018). *A practitioner's introduction to equating with primers on classical test theory and item response theory*. The Council of Chief State School Officers. https://ccsso.org/sites/default/files/201806/A%20Practitioners%20Introduction%20to%20Equating%20revised%20edition.pdf

Santos, C. M. C., Pimenta, C. A. M., & Nobre, M. R. C. (2007). The PICO strategy for the research question construction and evidence search. *Rev Latino-am Enfermagem, 15*(3), 508–5011. https://doi.org/10.1590/s0104-11692007000300023

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P): elaboration and explanation. *BMJ*, 349:g7647. https://doi.org/10.1136 /bmj.g7647

Shea, J. A., & and Norcini, J. J. (1995). *Licensure testing: Purposes, procedures, and practices*. University of Nebraska-Lincoln. https://digitalcommons.unl.edu/buroslicensure/16/

Sinharay, S. (2017). On the choice of anchor test in equating. *Educational Measurement: Issues and Practice, 37*(4), 1-6. https://doi.org/10.1111/emip.12175

Sinharay, S., Haberman, S., Holland, P., & Lewis, C. (2012). *A note on the choice of an anchor test in equating*. ETS. https://doi.org/10.1002/j.2333-8504.2012.tb02296.x

Sinharay, S., & Holland, P. W. (2006a). *The correlation between the scores of a test and an anchor test*. ETS. https://doi.org/10.1002/j.2333-8504.2006.tb02010.x

Sinharay, S., & Holland, P. W. (2006b). *Choice of anchor test in equating*. ETS. https://doi.org/10.1002/j.2333-8504.2006.tb02040.x

Sinharay, S., & Holland, P. W. (2007). Is it necessary to make anchor tests mini-versions of the tests being equated or can some restrictions be relaxed? *Educational Measurement, 44*, 249–275. https://doi.org/10.1111/j.1745-3984.2007.00037.x

Sirriyeh, R., Lawton, R., Gardner, P., & Armitage, G. (2011). Reviewing studies with diverse designs: the development and evaluation of a new tool. *Evaluation in Clinical Practice, 18*(4), 746-752. https://doi.org/10.1111/j.1365-2753.2011.01662.x

Suh, Y., Morch, A. A., Kane, M. T., & Ripkey, D. R. (2009). An empirical comparison of five linear equating methods for the NEAT design. *Measurement: Interdisciplinary Research and Perspectives, 7*(3), 147-173. https://doi.org/10.1080/15366360903418048

Sunnassee, D. (2011). *Conditions affecting the accuracy of classical equating methods for small sample under the NEAT design: A simulation study* (Doctoral Dissertation, University of North Carolina). https://libres.uncg.edu/ir/uncg/listing.aspx?id=8164

Tai, J., Ajjawi, R., Bearman, M., & Wiseman, P. (2020). Conceptualizations and measures of student engagement: A worked example of systematic review. In O. Zawacki-Richter, M. Kerres, S. Bendenlier, M. Bond & K. Buntins (Eds.), *Systematic reviews in educational research* (pp. 91-110). Springer. https://doi.org/10.1007/978-3-658-27602-7

Trierweiler, T. J., Lewis, C., & Smith, R. L. (2016). Further study of the choice of anchor tests in equating. *Educational Measurement*, *53*, 498–518. https://doi.org/10.1111/jedm.12128

van der Linden, W. J., & Wiberg, M. (2010). Local observed-score equating with anchor-test designs. *Applied Psychological Measurement, 34*(8), 620-640. https://doi.org/10.1177/0146621609349803

von Davier, A. A. (2008). New results on the linear equating methods for the non-equivalent-groups design. *Educational and Behavioral Statistics, 33*(2), 186-203. https://doi.org/10.3102/1076998608302633

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. Springer.

Wallin, G., Haggstrom, J., & Wiberg, M. (2021). How important is the choice of bandwidth in kernel equating? *Applied Psychology Measurement, 45*(7-8), 518-535. https://doi.org/10.1177/0146621621 1040486

Wei, H. (2010, May*). Impact of non-representative anchor items on scale stability* [Paper presentation]. National Council on Measurement in Education, Denver, Pearson.

Yang, W. L., & Houang, R. T. (1996, April). *The effect of anchor length and equating method on the accuracy of test equating: comparison of linear and IRT-based equating using an anchor-item design* [Paper presentation]. American Educational Research Association, New York, NY. https://eric.ed.gov/?id=ED401308

Yi, H. S. (2009). Evaluating the performance of non-equivalent groups anchor test equating under various conditions of anchor test construction. *Educational Evaluation, 22*(3), 847-869. https://www.kci.go.kr/kciportal/ci/sereArticleSearch/ciSereArtiView.kci?sereArticleSearchBean.artiId=ART001378603

Zhang, M., & Kolen, M. J. (2013). *Effect of the number of common items on equating precision and estimation of the lower bound to the number of common items needed.* Center for Advanced Studies in Measurement and Assessment (CASMA). https://www.education.uiowa.edu/casma