Educational Measurement and Evaluation Studies

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

National Organization for Educational Testing

A Comparison between Benchmarking and Bookmarking to Classification of Performance Levels in Large-scale Study of Mathematics Assessment

Masoud Kabiri[1]

1. Assistant Professor of Research Institute for Education; (Corresponding Author), E-mail: maskabiri@yahoo.comm

Article Info

Article Type:

Research Article

Received:

2021.04.11

Revised:

2021.11.30

Accepted:

2021.12.05

Published online:

2021.12.06

ABSTRACT

Objective: Standard setting is one of the assessment techniques to create valid classifications of examinees. In the present study, the effect of two standard setting methods, benchmark ing and bookmarking, was examined in results of a large-scale study, which was planned for assessing mathematics learning in sixth grade students of Tehran.

Methods: Two methods were compared using data of a provincial large-scale assessment which carried out on 9720 sixth grade students in Tehran. They asked 264 mathematics items and their response were analyzed by plausible values.

Results: Results of applying benchmark showed that 75, 48, 18, and 2 percent of students attained minimum scores in low, mediate, high, and advanced levels, respectively. In addition, 23.9 percent of items located in the same level that identified by content experts. In contrast, quality of classification by content experts in bookmarking was criticized due to comparing of successive averages with standard deviations of location parameters. Moreover, effect of using five response probabilities: 0.52, .057, 0.62, 0.67, and 0.75 in classification of students indicated that, in spite of recommendation of response probability 0.67 in literature, the lowest response probability (0.52) produced the most realistic results rather than other response probabilities, however, this is still a strictly standard comparing benchmarking methods.

Conclusion: Standard setting should be considered as a technical issue in all assessments that grading or pass/fail is consequent of the test.

Keywords: Standard setting, Benchmarking, Bookmarking, Math education.

How to Cite: Kabiri, Masoud (2021). A Comparison between Benchmarking and Bookmarking to Classification of Performance Levels in Large-scale Study of Mathematics Assessment. *Educational Measurement and Evaluation Studies*, 11 (34): 63-86 pages. DOI: 10.22034/EMES.2021.248192

Publisher: National Organization of Educational Testing (NOET)

مطالعات اندازه‌گیری و ارزشیابی آموزشی

سازمان سنجش آموزش کشور

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

مقایسهٔ روش‌های معیارگزینی نقطه‌گذاری معیار و علامت‌گذاری در دسته‌بندی سطوح عملکرد مطالعه کلان‌مقیاس سنجش ریاضی

مسعود کبیری[1]

۱.استادیار پژوهشگاه مطالعات آموزش‌وپرورش، تهران، ایران؛ (نویسنده مسئول)، پست الکترونیک:maskabiri@yahoo.com

اطلاعات مقاله

نوع مقاله:
مقاله پژوهشی

دریافت:
۱۴۰۰/۰۱/۲۲

اصلاح:
۱۴۰۰/۰۹/۰۹

پذیرش:
۱۴۰۰/۰۹/۱۴

انتشار:
۱۴۰۰/۰۹/۱۵

چکیده

هدف: معیارگزینی یکی از فنون سنجش برای طبقه‌بندی معتبر آزمودنی‌ها است. در این مطالعه، تأثیر استفاده از دو روش معیارگزینیِ نقطه‌گذاری معیار و علامت‌گذاری بر نتایج حاصله از مطالعه کلان‌مقیاسی تحلیل شد که برای سنجَش یادگیری ریاضی پایه ششم در بین دانش‌آموزان شهر تهران اجرا شده بود.

روش پژوهش: این روش‌ها روی داده‌های سنجش کلان‌مقیاس استانی که بر ۹۷۲۰ دانش‌آموز پایهٔ ششم شهر تهران اجرا شده بود، مقایسه شدند. مشارکت‌کنندگان در این پیمایش درمجموع ۲۶۴ سؤال ریاضی را پاسخ دادند و پاسخ‌های آنان با استفاده از روش مقادیر محتمل تحلیل شدند.

یافته‌ها: نتایج نشان دادند که به‌کارگیری روش نقطه‌گذاری معیار باعث می‌شود که به ترتیب ۷۵، ۴۸، ۱۸ و ۲ درصد از دانش‌آموزان حداقل نمرات لازم را در سطوح عملکردی پایین، متوسط، بالا و پیشرفته کسب کنند. همچنین، با استفاده از این روش ۲۳/۹ درصد از سؤال‌ها در همان سطحی قرار گرفتند که توسط کارشناسان موضوعی تعیین شده بودند. در مقابل، مقایسه فاصله میانگین‌های متوالیِ پارامتر جایگاه با انحراف معیار جایگاه در سطوح عملکردی، کیفیت دسته‌بندی اولیهٔ کارشناسان برای استفاده در روش علامت‌گذاری را زیر سؤال برد. افزون بر این، تأثیر استفاده از پنج احتمال پاسخ ۰/۵۲، ۰/۵۷، ۰/۶۲، ۰/۶۷ و ۰/۷۵ بر دسته‌بندی دانش‌آموزان نشان داد که با وجود تأکید پیشینه پژوهشی بر احتمال پاسخ ۰/۶۷، کمترین احتمال پاسخ (۰/۵۲) نتایج واقعی‌تری را نسبت به بقیه تولید می‌کند ولی همچنان در مقایسه با روش نقطه‌گذاری معیار، معیار سخت‌گیرانه‌ای به نظر می‌رسد.

نتیجه‌گیری: باید به معیارگزینی به‌عنوان یک مبحث فنی در همه سنجش‌هایی که درجه‌بندی یا قبول و ردی یکی از پیامدهای شرکت در آزمون است، توجه بیشتری شود.

واژگان کلیدی: معیارگزینی، نقطه‌گذاری معیار، علامت‌گذاری، آموزش ریاضی

استناد: کبیری، مسعود (۱۴۰۰). مقایسهٔ روش‌های معیارگزینی نقطه‌گذاری معیار و علامت‌گذاری در دسته‌بندی سطوح عملکرد مطالعه کلان‌مقیاس سنجش ریاضی. فصلنامه مطالعات اندازه‌گیری و ارزشیابی آموزشی، ۱۱(۳۴)، ۸۶-۶۳ صفحه،
DOI: 10.22034/EMES.2021.248192

مقدمه

با اینکه هدف اولیهٔ آزمون‌ها تعیین نمره برای توانایی افراد است، ولی در نگاه جامع‌تر، اطلاع از نمره به‌خودی‌خود همه کاربردهای آزمون را در برنمی‌گیرد. به‌عنوان مثال، هدف آزمون‌های اعطای گواهی، که برای تعیین کفایت یادگیری طرح‌ریزی می‌شـــود، تعیین این اســـت که آیا آزمودنی صلاحیت کافی برای قبولی یا ورود به مرحله بعدیِ شـــغلی یا تحصیلی را دارد یا خیر. بسیاری از آزمون‌های معمول در مؤسسات آموزشی نیز بدین منظور طرح‌ریزی می‌شـــود. بااین‌حال، همچنان تردید جدی در مورد مناسب بودن معیار قبولی به دلیل تأثیرگذاری آن بر تصمیمات بعدی وجود دارد. افزون بر این، دسته‌بندی آزمودنی‌ها به سطوح عملکردی مختلف حتی در مطالعاتی که کمتر در معرض تبعات تصمیمات هستند (همچون آزمون‌های غیر سرنوشت‌ساز) نیز جذاب است، زیـــرا اطلاعات عملکرد آزمودنی‌ها را بهتر می‌تواند برای سیاســـت‌گذاران خلاصه کند. احتمالاً به همین دلیل مؤسسه آمار یونسکو[1] به همه کشورها در برنامه اتحاد جهانی برای نظارت بر یادگیری[2] توصیه کرده است که برای تسهیل اندازه‌گیری و گزارش خروجیِ میزان یادگیری، خلاصه سطوح یادگیری را منتشر کنند (مؤسسه آمار یونسکو، ۲۰۱۷). فرایندی که در خلاصه‌سازی یادگیری بر اساس سطوح عملکرد مورد استفاده قرار می‌گیرد، به‌عنوان معیارگزینی[3] شناخته می‌شود.

معیارگزینی به‌عنوان فرایند در نظر گرفتنِ معیارهای عملکردی معنی‌دار تعریف شده است (کارترایت[4]، ۲۰۱۵). برای درک ساده آنچه در فرایند معیارگزینی اتفاق می‌افتد، دو توصیف مفید است: تعریف مجموعه‌ای از نقاط برش در مقیاس توانایی، یکی از این توصیف‌ها است که در هر فاصله نقاط، شایستگی‌های خاصی از آزمودنی‌ها مورد انتظار اســـت. ازآنجاکه مقیاس توانایی پیوســـته بوده و هیچ نقطه برش طبیعی ندارد، عمل تقسیم‌بندی مقیاس توانایی به ســـطوح توانایی بر اســـاس نقاط دلخواهی (سازمان توســـعه و همکاری اقتصادی[5]، ۲۰۱۷) صورت می‌پذیرد. بنابراین، تصمیم‌گیری درباره تعداد سطوح و تعیین نقاط برش بر اساس ترکیبی از قضاوت‌ها، ملاحظات روان‌سنجی و کاربردپذیری نقاط برشِ به‌کاررفته (پرایس[6]، ۲۰۱۷) موضوعی است که در معیارگزینی پی گرفته می‌شود. در توصیف دوم نیز می‌توان معیارگزینی را به خوشه‌بندی سؤال‌ها در مجموعه‌هایی تعریف کرد که سطوح مختلف عملکرد را اندازه‌گیری می‌کنند (لیسیتز[7]، ۲۰۱۳). هر کدام از این دو توصیف می‌توانند برای درک ماهیت معیارگزینی مفید باشند. به‌طورکلی، اقدامات و رویه‌های معیارگزینی شامل تبدیل مقیاس توانایی پیوسته به چند مقوله است که درجات عملکرد را بهتر نشان می‌دهد.

برای بالندگی مفهوم و روش‌های معیارگزینی بسترهایی مؤثر بوده‌اند که اولین آن به رویه‌های اجرایی پس از

1. UNESCO Institute for Statistics (UIS)
2. Global Alliance to Monitor Learning
3. standard setting
4. Cartwright
5. The Organisation for Economic Co-operation and Development (OECD)
6. Price
7. Lissitz

قانون «هیچ دانش‌آموزی نباید عقب بیافتد[1]» در آمریکا مربوط بود. در این قانون تصریح شده بود که هر ایالت، سطوح عملکردی مورد انتظار خود را تعیین کند و درصد دانش‌آموزان در هر یک از سطوح را به‌طور مرتب گزارش دهد. در کنار این موضوع، رویگردانی متخصصان از آزمون‌ها و تحلیل‌هایی که مبتنی بر تحلیل‌های هنجار مرجع[2] بودند را نیز می‌توان اضافه کرد. این‌گونه تحلیل‌ها که بر اساس رتبه‌بندی پایه‌ریزی شده بود برای بسیاری از تصمیمات مهم همچون ارتقا، گواهی دهی، ابقا، جایگزینی و ... ناکارآمد بودند و به‌تدریج توجهات به آزمون‌ها و تحلیل‌هایی متمرکز شد که مستلزم اتمام کار و بر اساس آزمون‌های ملاک مرجع[3] و معیار مرجع[4] پایه‌ریزی شده بود (سیزک و بلانچ[5]، ۲۰۰۷). مجموعه این تحولات، نظامی منطقی از قواعد و رویه‌ها را برای پاسخگویی به این نیازها ایجاب کرد.

در بافت آموزش، زمانی که از معیار[6] صحبت می‌کنیم، دو مفهوم مورد توجه هستند: اول، معیارها به‌عنوان انتظارات در نظر گرفته شده که در برنامه‌های درسی رسمی، به‌صورت هدفِ قصدشده بیان می‌شوند؛ مفهوم دوم به ملاک‌های کیفیت توافق‌شده برای موضوعات خاص مثل معیارهای تدریس یا معیارهای سنجش مربوط می‌شود. وقتی از معیارگزینی صحبت می‌کنیم، هر دو مفهوم فوق مد نظر ماست (السون و نیلسون[7]، ۲۰۱۷)؛ به‌طوری‌که اگر توصیه اولیهٔ معیار عملکردی را در نظر بگیریم، هر معیار عملکردی، نقطه برشی است که انتهای بالاییِ سطح عملکردی را محدود می‌کند (فیلیپس[8]، ۲۰۱۲). علاوه بر مفهوم قبلی، مفهوم آزمودنی خط مرزی[9] (آزمودنیِ به‌طور حداقلی قابل قبول) نیز مفهوم بسیار مهمی است که فصل مشترک همه روش‌های معیارگزینی به شمار می‌رود و به معنای آزمودنیِ فرضی است که بر اساس قضاوت کارشناسان موضوعی می‌تواند به سؤال‌های سطح توانایی مشخصی پاسخ دهد (پرایس، ۲۰۱۷). هر سطح توانایی از طریق توصیفگرهای سطوح عملکردی[10] تعریف می‌شود تا تفاوت اصلی در سطوح افزایشی عملکرد را به‌خوبی تعریف کند. هر یک از این توصیفگرها مشخص می‌کند که در هر سطح، چه انتظاری از دانش‌آموزان برحسب دانش و مهارت وجود دارد.

روش‌های زیادی برای اجرای معیارگزینی پیشنهاد شده است. روش‌های سنتی بر اساس قضاوت‌های فردی بودند که با برخی از آماره‌ها و شاخص‌های روان‌سنجی همراه شده بودند و از روش‌های تحلیل کلاسیک آزمون‌سازی بهره می‌بردند. ازجمله این روش‌ها می‌توان به روش ندلسکی[11] و روش ایبل[12] اشاره کرد. معروف‌ترین روش در

1. No Child Left Behind Act
2. norm-reference tests
3. criterion reference tests
4. standard-reference tests
5. Cizek & Bunch
6. standard
7. Olsen & Nilsen
8. Phillips
9. borderline examinee
10. performance level descriptions
11. Nedelsky method
12. Eble method

این دســـته روش انگاف[1] است. در روش انگاف از متخصصان موضوعی خواسته می‌شود که محتوای سؤال‌های آزمون را بررســـی کرده و در مورد نسبتی از آزمودنی‌های جامعه هدف که به سؤال پاسخ خواهند داد، قضاوت کنند. این فرایند برای هر سؤال تکرار شده و مجموع نمره‌های مورد انتظار برای همه سؤال‌ها، نمره آزمودنی‌ای که به‌طور حداقلی قابل قبول است را نشان می‌دهد. عملیاتی کردن «آزمودنی با کمینهٔ ملاکِ قابل قبول» یکی از فعالیت‌های اصلی در این روش اســـت که این مرحله با کمک توصیفگرهای سطوح عملکردی اجرا می‌شود. روش انگاف به‌طور اصلی برای سؤال‌های دوبخشی مثل سؤال‌های چندگزینه‌ای و بازپاسخ با نمره‌های درست یا نادرست طراحی شده بود. ولی برای پوشش سؤال‌های بازپاسخِ چندبخشی، روش اصلاح‌شده انگاف معرفی شد. در این روش از متخصصان موضوعی خواسته می‌شود در سؤال‌های بازپاسخِ چندبخشی، نمره‌ای را تعیین کنند که «آزمودنی با کمینهٔ ملاکِ قابل قبول» از بین طیف نمره‌های هر ســـؤال دریافت خواهد کرد (پرایس، ۲۰۱۷؛ لیسیتز، ۲۰۱۳). میانگین نمره‌های همه متخصصان و تقسیم آن بر تعداد نمره‌های ممکن در هر طیف، احتمال پاســـخ به سؤال را در همان مقیاس سؤال‌های دوبخشی قرار می‌دهد. به‌طورکلی، روش انگاف به دلیل سادگی مورد اقبال بسیاری آزمون‌سازان و پژوهشگران است. بااین‌حال، به دلیل عدم برخورداری از آزمون‌های مقدماتی و داده‌های حاصل جهت شکل دادن به قضاوت متخصصان و جلوگیری از برآوردهای غیر واقعی آنان مورد نقد قرار گرفته است (پرایس، ۲۰۱۷). به این معنا که عموم روش‌های سنتی که مبتنی بر محتوا هستند به‌طور زیادی متکی بر افراد هستند. تعیین معیارهای دقیق بدون مرجع بیرونی و صرفاً با تکیه بر تجارب کلاسی توســـط فیلیپس (۲۰۱۲) به پرواز بدون رادار تعبیر شـــده است و می‌تواند به اثر دریاچه ووبگان[2] (وضعیتی که پیشرفت تحصیلی اکثر دانش‌آموزان بیش‌برآورد و بیش از میانگین برآورد می‌شود) منجر شود. استیلای چند دهه‌ایِ روش انگاف بر ادبیات معیارگزینی با معرفی روش علامت‌گذاری[3] در انتهای دهه ۹۰ میلادی با تلاش پژوهشگران دانشگاه مک‌گرو هیل[4] کانادا به پایان رسید (سیزک و بانچ، ۲۰۰۷). روش علامت‌گذاری با اتکا به تحلیل‌های مبتنی بر نظریه پرسش پاسخ[5]، مجموعه کاملی از فعالیت‌ها برای تعیین نقاط برش را مطرح کرد. به دلیل نتایج قابل اعتماد حاصله و همچنین سادگی استفاده از این روش نسبت به سایر روش‌های معیارگزینی کاربرد آن به‌طور قابل ملاحظه‌ای متداول شـــد؛ به‌طوری‌که در ســـنجش‌های ملی مهم همچون سنجش ملی پیشرفت آموزشی[6] در آمریکا از سال ۲۰۰۵ به‌کاررفته است. همچنین، این روش در سنجش‌های ملی کشورهایی چون پرو، بلژیک، برمودا، قطر، کره جنوبی و اســـرائیل نیز به‌کاررفته است (لوئیز و همکاران[7]، ۲۰۱۲). انتخاب نام این روش بر اساس شیوه کاری است که از متخصصان موضوعی انتظار می‌رود؛ به‌طوری‌که از آنان خواسته

1. Angoff method
2. Lake Wobegon Effect
3. bookmark
4. McGraw Hill University
5. Item Response Theory (IRT)
6. National Assessment Educational Progress (NAEP)
7. Lewis et al

می‌شود قضاوت خود را با علامت‌گذاری در دفترچه حاوی سؤال‌های آزمون نشان دهند.

روش علامت‌گذاری با تعریف توصیفگرهای سطوح عملکردی برای همه سطوح عملکردی شروع می‌شود. فرض استفاده از این روش آن است که سؤال‌ها قبلاً در بین آزمودنی‌ها اجرا شده باشند. چنانچه نتایج تحلیل داده‌ها در دسترس باشد، می‌توان سؤال‌ها را بر اساس پارامتر دشواری سؤال‌ها (بر مبنای نظریه پرسش پاسخ) از ساده به دشوار مرتب کرد. مرتب کردن سؤال‌ها به معنای آن است که هر چه سؤال‌ها جلوتر می‌رود، توانایی‌های شناختی بیشتری برای پاسخ درست به سؤال‌ها مورد نیاز است. انجام این کار در سؤال‌های دوبخشی روشن است، ولی در سؤال‌ها با نمره‌های چندبخشی، هر کدام از بخش‌های اضافیِ سؤال به‌عنوان یک سؤال مجزا در نظر گرفته شده و به‌طور جداگانه مرتب می‌شود. به‌عنوان مثال، سؤال ۲ نمره‌ای با دامنه نمره‌های ۰، ۱ و ۲، به‌صورت دو سؤال مجزایِ نمره کامل (نمره ۲) و نمره ناقص (نمره ۱) در بین سؤال‌ها مرتب می‌شود.

تولید دفترچه سؤال مرتب‌شده به‌منظور ارائه به متخصصان موضوعی است تا محتوای خوشه‌های کوچک‌ترِ سؤال‌ها را بررسی کنند. از این گروه خواسته می‌شود تا سؤالی را در مجموعه سؤال‌های مرتب‌شده مشخص کنند که انتظار می‌رود آزمودنی‌های خط‌مرزی به محتوای سؤال‌های قبل از آن مسلط خواهند بود. برای این کار، متخصصان موضوعی بحث می‌کنند که چه مواردی سؤال یا گروهی از سؤال‌ها را دشوارتر کرده و بر اساس خواسته‌های شناختی سؤال، چه نقطه‌ای می‌تواند برای علامت‌گذاریِ متناظر با هر سطح عملکردی بهترین انتخاب باشد. تفاوت کار متخصصان در روش انگاف و علامت‌گذاری در این است که در روش انگاف از افراد خواسته می‌شود که قضاوت‌های احتمالی را برای هر سؤال ارائه دهند ولی در روش علامت‌گذاری قضاوت‌ها محدود به تعداد سطوح عملکردی منهای یک (آستانه‌های سطوح عملکردیِ متوالی) خواهد بود. پرسش اصلی که از متخصصان موضوعی پرسیده می‌شود این است که آیا محتمل است که آزمودنی با حداقل توانایی در سطح عملکردی مشخصی به سؤال پاسخ دهد؟ تعریف یا عملیاتی‌سازی قاعده تصمیم‌گیریِ محتمل بودن در سؤال فوق یکی از موضوعات مهم در روش علامت‌گذاری است که با عنوان احتمال پاسخ[1] شناخته می‌شود. احتمال پاسخ، احتمالی است که آزمودنی در میانه سطح عملکردی به‌طور درستی به سؤال‌ها با دشواری متوسط برای همان سطح پاسخ دهد (سازمان توسعه و همکاری اقتصادی، ۲۰۱۷). احتمال پاسخ برای این استدلال، منطقی فراهم می‌کند که آزمودنی‌های نزدیک به جایگاه سؤال[2] در مقیاس، خصیصه‌های متناظر برای تسلط و پاسخگویی را دارا هستند (لوئیز و همکاران، ۲۰۱۲).

احتمال پاسخ، معمولاً بین ۰/۵۰ تا ۰/۸۰ در نظر گرفته می‌شود. مقادیر زیر ۰/۵۰ به دلیل اینکه با مفهوم تسلط هم‌خوان نیستند، به‌کاربرده نمی‌شوند زیرا احتمال کمی را در نظر می‌گیرند. از سوی دیگر، احتمال پاسخ ۰/۸۰ نیز خیلی بالا است. درنتیجه فاصله بین این دو احتمال می‌تواند مقدار مطلوبی باشد. به‌عنوان مثال، در مطالعه

1. response probability
2. item location

پیزا[1] در سال ۲۰۰۰ احتمال پاسخ ۰/۶۲ در نظر گرفته شد (سازمان توسعه و همکاری اقتصادی، ۲۰۱۷). بااین‌حال، مطالعات نشان می‌دهد که احتمال پاسخ ۰/۶۷ از لحاظ آماری بهینه بوده و بیشترین اطلاعات را برای ایجاد قاعده‌های تصمیم‌گیری به وجود می‌آورد (سیزک و بانچ، ۲۰۰۷؛ کارترایت، ۲۰۱۵)؛ به‌طوری‌که بررسی داده‌های تجربی نشان داد توزیع گویه‌ها در مقیاس، با این میزان احتمال پاسخ با دامنه وسیعی از سطوح توانایی مناسبت دارد. علاوه‌براین، احتمال پاسخ ۰/۶۷ برای قضاوت با مفهوم تسلط مرتبط‌تر است و توسط متخصصان راحت‌تر درک می‌شود زیرا می‌توان آن را به‌سادگی به شکل دوسوم عملیاتی کرد (لوئیز و همکاران، ۲۰۱۲).

پس از مرتب کردن سؤال‌ها برحسب احتمال پاسخ و ارائه به متخصصان موضوعی، سؤال‌های معرف هر یک از سطوح عملکردی بر اساس میانگین نظر متخصصان مشخص می‌شوند. در حقیقت، فواصلی بین دو سؤال مشخص می‌شود که احتمال داده می‌شود آزمودنی‌های معرف سطح عملکردیِ پایین‌تر نمی‌توانند به سؤال‌های بعد از نقطه تعیین شده پاسخ دهند. سطح متناظر با دشواری این نقطه در مقیاس توانایی که مرتبط با احتمال پاسخِ تعیین شده (به‌طور مثال ۰/۶۷) است را به‌عنوان نقطه برش بین دو سطح عملکردیِ متوالی در نظر می‌گیرند. برای ارزشیابی اینکه تا چه اندازه سطوح به‌خوبی تعریف شده‌اند، انحراف معیارِ دشواری سؤال‌ها بررسی می‌شود. درصورتی‌که انحراف معیار دشواریِ سؤال‌ها در سطح عملکردی بیش از فاصله بین میانگین‌ها یا آستانه‌های سطوح متوالی باشد، تعریف سطوح از لحاظ آماری ضعیف ارزیابی می‌شود (کارترایت، ۲۰۱۵).

در انتخاب نقاط برش، ملاحظات دیگری را هم می‌توان در نظر داشت. اول آنکه اگر آزمودنی‌های درون یک سطح بسیار کم باشند، هر گونه آماره‌های مرتبط با این گروهِ، بسیار کوچک، بی‌ثبات و غیر قابل تفسیر خواهند بود (کارترایت، ۲۰۱۵). همچنین، بدون آنکه هیچ دلیل ریاضی وجود داشته باشد معمولاً در سنجش‌های ملی و بین‌المللی آستانه‌هایی با فواصل مساوی رعایت می‌شود تا ارتباط مستقیم‌تری با مخاطبان پیدا کند. بااین‌حال، سازمان همکاری و توسعه اقتصادی (۲۰۱۷) تفاوت‌های جزئی بین سطوح را مجاز می‌داند.

روش علامت‌گذاری به چندین دلیل مورد استقبال افرادی قرار گرفت که مایل به اجرای معیارگزینی بودند: اول اینکه در این روش هر دو نوع سؤالِ چندگزینه‌ای و پاسخ‌سازِ چندبخشی[2] قابل پوشش هستند. هرچند این قابلیت برای روش انگافِ اصلاح‌شده هم وجود دارد ولی ارزیابی می‌شود که خواسته‌های شناختی از متخصصان، هنگام بررسی سؤال‌های بازپاسخِ چندبخشی بیشتر از سؤال‌های پاسخ‌گزینِ دوبخشی است و درنتیجه کارایی روش انگاف در هنگام مواجهه با آزمون‌های ترکیب‌شده از این دو نوع سؤال کاهش می‌یابد (لوئیز و همکاران، ۲۰۱۲)؛ دلیل دوم محبوبیت روش علامت‌گذاری این است که از نقطه‌نظر متخصصان موضوعی کار نسبتاً ساده‌تری به آنان واگذار می‌شود زیرا به جای تحلیل در سطح تک‌تک سؤال‌ها ، تنها در سطوح عملکردی به قضاوت می‌پردازند؛ سومین دلیل به بهره‌مندی روش علامت‌گذاری از مزایای نظریه پرسش پاسخ مربوط می‌شود

1. Programme for International Student Assessment (PISA)
2. created-response polytomous items

(سیزک و بانچ، ۲۰۰۷). مهم‌ترین ویژگی نظریه پرسش پاسخ در معیارگزینی این است که دشواری سؤال‌ها و توانایی افراد در یک مقیاس مشترک قرار می‌گیرند و امکان مقایسه با یکدیگر ایجاد می‌شود. درنهایت، می‌توان به سهولت تعریف چندین نقطه برش و درنتیجه تولید سطوح عملکردی متعدد در این روش اشاره کرد. در کنار این مزایا، استلزام برخورداری از داده‌های تجربی برای استفاده از این روش را نیز می‌توان به‌عنوان محدودیت این روش برشمرد (کارترایت، ۲۰۱۵).

روش دیگر معیارگزینی که امروزه در مطالعات کلان‌مقیاس بیشـــتر کاربرد دارد، با عنوان نقطه‌گذاری معیار[1] شـــناخته می‌شود. این روش امروزه در مطالعات تیمز، پرلز و پیزا به کار می‌رود که در دو مطالعه اول با عنوان لنگرگزینی[2] نیز شـــناخته می‌شـــود. چند ویژگی، روش نقطه‌گذاری معیار را از ســـایر روش‌های معیارگزینی به‌خصوص روش علامت‌گذاری متمایز می‌کند. اول اینکه داده‌های مرجعِ بیرونی نقطه شروع و نه پایان روش نقطه‌گذاری معیار به شمار می‌رود. استدلال پشتیبان این نگاه آن است که معیارهای عملکردی اساساً تصمیمات سیاست‌گذاری محسوب می‌شوند تا تصمیمات محتوایی، و باید با دانش جهان واقعی و ملزومات مرتبط با آن هدایت شـــود. دومین ویژگی به تطبیقی بودن این روش مربوط می‌شـــود؛ به این شـــکل که اگرچه این روش معیارگزینی را نمی‌توان یک رویکرد هنجاری به شمار آورد، ولی هنجارهای ملی و بین‌المللی بیرونی می‌توانند راهنمای خوبی برای دانســـتن این موضوع باشـــند که دانش‌آموزان چه می‌دانند و چه می‌توانند انجام دهند. بنابراین، هدف روش نقطه‌گذاری معیار از این منظر، تعیین سطح دانش و مهارت در رقابت با وضعیت ملی یا بین‌المللی است (فیلیپس، ۲۰۱۲). به همین دلیل، پیوند[3] نتایج حاصله با مطالعات ملی یا بین‌المللی یکی از مراحل این روش را تشکیل می‌دهد.

شـــیوهٔ کار روش نقطه‌گذاری معیار را می‌توان با بررســـی روش کار در مطالعات تیمز و پرلز توضیح داد. بدین منظور، نخست داده‌های مربوط به هر سؤال و هر آزمودنی با استفاده از نظریه پرسش پاسخ تحلیل می‌شوند و تصمیماتی در مورد تعداد و محل هر یک از نقاط برش اتخاذ می‌شود. سپس، آزمودنی‌هایی که در محدوده ۵ نمره‌ای هر سطح عملکردی (در مقیاس با میانگین ۵۰۰ و انحراف معیار ۱۰۰) قرار گرفته‌اند، برای تحلیل‌های بعدی مشـــخص می‌شوند. این محدوده ۱۰ نمره‌ای هم نمونه مناســـبی از نمره‌های دانش‌آموزان در هر سطح عملکردی به حســـاب آمده و هم برای متمایز شـــدن عملکرد در یک سطح عملکردی نسبت به سطوح دیگر مناسب است. در گام بعدی، درصد پاسخگویی این آزمودنی‌ها در محدوده هر یک از سطوح عملکردی با استفاده از وزن نمونه‌گیری مناسب محاســـبه می‌شود. این محاسبه برای سؤال‌های چندگزینه‌ای و بازپاسخِ دوبخشی مشـــخص است و برای سؤال‌های بازپاسخِ چندبخشی به تفکیک برای هر بخش محاسبه می‌شود. پس از این

1. benchmarking
2. anchoring
3. linking

گام، ملاک‌هایی برای تعیین سؤال‌های متناظر با هر سطح عملکردی، به کار برده می‌شوند. بر اساس محتوا یا توصیفات سؤال‌های درون هر سطح، توصیفگرهای سطوح عملکردی به‌گونه‌ای تدوین می‌شود که عملکرد هر فردِ قرار گرفته شـــده در آن سطح یا سطوح بالاتر را به‌روشنی توصیف کند (السن و نیلسن، ۲۰۱۷؛ مولیس و همکاران[1]، ۲۰۱۶). ملاک‌هایی که برای بررسی سؤال‌ها به کار می‌رود بدین شرح هستند:

در سؤال‌های چندگزینه‌ای، حداقل ۶۵ درصد از آزمودنی‌های قرار گرفته در همان سطح عملکردی و کمتر از ۵۰ درصد از آزمودنی‌های قرار گرفته در ســـطح پایین‌تر، پاسخ درست داده باشند. بخش دوم این ملاک برای اولین سطح عملکردی مورد توجه قرار نمی‌گیرد.

در سؤال‌های بازپاسخ، حداقل ۵۰ درصد از آزمودنی‌های قرار گرفته شده در همان سطح عملکردی پاسخ درست داده باشند. برای این‌گونه سؤال‌های بخش دوم ملاک فوق اعمال نمی‌شود.

تفاوت بین مقادیر ملاک‌ها در دو نوع سؤال به دلیل وجود حدس در سؤال‌های چندگزینه‌ای است. برای اینکه دامنه سؤال‌ها برای پوشش بهتر توصیفگرهای سطوح عملکردی بیشتر شود، مفهوم «تقریباً معیارگزین شده[2]» معرفی شده است. سؤال‌های چندگزینه‌ای که بین ۶۰ تا ۶۵ درصد از آزمودنی‌ها در یک سطح عملکردی به آن پاسخ درست داده باشند را تقریباً معیارگزین شده می‌نامند. بخش دوم ملاک که مربوط به میزان پاسخِ سطح عملکردیِ قبل از خود است در این‌گونه سؤال‌ها اعمال نمی‌شود. علاوه‌براین، سؤال‌های چندگزینه‌ای که کمتر از ۶۰ درصد و سؤال‌های بازپاسخ که کمتر از ۵۰ درصد از آزمودنی‌های بالاترین سطح عملکرد به آن پاسخ درست داده‌اند، به‌عنوان سؤال‌های خیلی دشوار طبقه‌بندی می‌شوند (مولیس و همکاران، ۲۰۱۶).

چنانچه ملاحظه می‌شـــود در روش نقطه‌گذاری معیار، انتخاب نقطه برش بر عهده گروه متخصصان نیســـت. این موضوع از آنجا ریشـــه می‌گیرد که اعتقادی به ایجاد سطوح عملکردی با تکیه صرف بر نظریه وجود ندارد بلکه تصمیمات در مورد تعداد و محل نقاط برش بر اســـاس ترکیبی از ملاک‌های عملیِ مرتبط با سودمندی و ملاک‌های تجربی اســـت (السن و نیلسن، ۲۰۱۷). در این روش انتخاب نقطه برش، برخلاف سایر روش‌های معیارگزینی، جزو گام‌های اولیهٔ معیارگزینی به شمار می‌رود.

با وجود مشـــابهت مطالعات کلان‌مقیاس در استفاده از روش نقطه‌گذاری معیار، بین شیوه کار مطالعه تیمز و پیزا تفاوت‌هایی نیز وجود دارد. درحالی‌که مطالعه تیمز، نمره‌های برش را در مقیاس بر اســـاس نقاط از پیش تعریف شـــده در نظر می‌گیرد، در مطالعه پیزا ملاک‌های دیگری برای انتخاب نقاط برش توجه می‌شود که به فاصله‌های نسبتاً نامنظم می‌انجامد. بنابراین، جایگاه دقیق نقاط برش در موضوعاتِ مورد سنجش در مطالعه پیزا مساوی نیست؛ درحالی‌که در مطالعه تیمز درس‌های مختلف نقاط برش یکسانی دارند. با وجود این، فاصله بین دو نقطه برشِ متوالی ۰/۷۵ واحد انحراف معیار در هر دو مطالعه اســـت (السن و نیلسن، ۲۰۱۷). چنانچه قرار

1. Mullis et al
2. almost anchored

باشد مجموعه‌ای از معیارها برای درس‌های مختلف تولید شود این موضوع چالش جدی به شمار می‌رود، زیرا معیارهای عملکردی باید به‌طور منطقی بین درس‌ها و بین پایه‌ها نسبتاً متجانس باشند. این موضوع با عنوان تنظیم‌بندی[1] شناخته می‌شود (فیلیپس، ۲۰۱۲).

با اینکه روش‌های مختلف به نتایج مختلفی منجر می‌شوند، تاکنون مشخص نشده است که روش‌های خاصی همیشه بر روش‌های دیگر برتری دارند، بلکه احتمالاً برخی از روش‌ها در انواع خاصی از آزمون‌ها یا شرایط بهتر از سایرین هستند. همچنین، توصیه‌های اندکی در مورد قواعد مشخصِ استفاده از روش‌ها در شرایط خاص ارائه شده است (سیزک و بانچ، ۲۰۰۷).

با وجود اهمیت به‌کارگیری روش‌های معیارگزینی، اســـتفاده از آنها در کاربردهای آزمون‌ســـازی داخل کشور چندان رایج نیست. مرور پیشینه در این زمینه چند کاربرد انگشت‌شمار را نشان می‌دهد که بیشتر مربوط به آموزش پزشکی اســـت. آزمون بالینی ساختارمند عینی[2] (مکارم و همکاران، ۱۳۹۶؛ جلیلی و مرتاض هجری، ۱۳۹۱؛ مرتـــاض هجری و همکاران، ۱۳۹۰) و آزمون پیش‌کارورزی (پره‌انترنی) علوم پزشـــکی (حبیب‌زاده و همکاران، ۱۳۹۸) در حوزه پزشکی و آزمون زبان انگلیسی (MSRT) (جلالی‌زاده و همکاران، ۱۳۹۸) آزمون‌هایی هستند که بر روی آنها معیارگزینی انجام گرفته است. با اینکه روش انگاف بیش از روش‌های دیگر توسط این پژوهشگران استفاده شده بود (جلالی‌زاده و همکاران، ۱۳۹۸؛ جلیلی و مرتاض هجری، ۱۳۹۱؛ مرتاض هجری و همکاران، ۱۳۹۰)، ولی از روش‌های دیگری همچون مدل راش (روش نقشه سؤال)[3]،کوهن[4]، رگرسیون مرزی[5]، هافستی[6] و علامت‌گذاری نیز استفاده شده است. این مطالعات که همگی حول تعیین یک نقطه برش بهینه برای تقسیم طیف به دو قسمت قبول یا رد تنظیم شده بودند، نشان دادند که نقاط برش انتخاب‌شده با ملاک‌های اولیهٔ تفاوت دارد. به‌طور مثال، جلالی‌زاده و همکاران (۱۳۹۸) نقطه برش آزمون MSRT را با استفاده از روش انگاف ۵۳/۶۶ و با روش علامت‌گذاری ۵۴/۲۷ تعیین کردند که با نقطه برش سنتی این آزمون (۵۰) متفاوت بود. افزون بر کاربرد روش‌های معیارگزینی در آزمون‌های سرنوشت‌ساز، این روش‌ها در آزمون‌های غیر سرنوشت‌ساز از جنس مطالعات کلان‌مقیاس نیز بسیار مهم هستند. به دلیل اینکه مطالعات کلان‌مقیاس در ایران چندان متداول نشده است، طبیعی است که روش‌های معیارگزینی برای این منظور نیز مهجور مانده است. بااین‌حال، باید توجه داشت که روش‌هایی از معیارگزینی در مطالعات کلان‌مقیاس کاربرد دارند که اولاً در گستره‌ای از انواع سؤال‌ها (شامل چندگزینه‌ای، کوتاه پاسخ و بازپاسخِ چندنمره‌ای) به‌سهولت قابل کاربرد بوده، ثانیاً به دلیل عدم وابستگی نتایج به نمونه، مبتنی بر نظریه پرسش پاسخ باشند و ثالثاً امکان تولید چندین نقطه برش و درنتیجه چندین

1. articulation
2. objective structured clinical examination (OSCE)
3. Rasch (item map)
4. Cohen
5. borderline regression
6. Hofstee

سطح عملکردی وجود داشته باشد. مرور روش‌های معیارگزینی نشان می‌دهد که چنین قابلیتی در دو روش نقطه‌گذاری معیار و علامت‌گذاری وجود دارد که اتفاقاً در بیشتر مطالعات کلان‌مقیاس کاربرد دارند. بنابراین، پرسش اصلی این خواهد بود که استفاده از هر یک از این روش‌ها چه تأثیری بر نتایج بررسی سطوح عملکردیِ مطالعه «برنامه رصد کیفیت آموزشی شـــهر تهران» (برکات) در بخش سنجش عملکرد ریاضی دانش‌آموزان دارد. این مطالعه توســـط اداره کل آموزش‌وپرورش شهر تهران با هدف پایش کیفیت آموزشی و فراهم کردن داده‌های باکیفیت برای پشتیبانی از توسعه سیاست‌گذاری آموزشی طراحی شده است. این برنامه، اولین مطالعه کلان‌مقیاس سنجش کیفیت یادگیری در سطح استانی اســـت که در کشور برگزار می‌شود (کبیری، ۲۰۱۹، ۱۳۹۹) که دو موضوع ریاضی (پایه ششم) و مهارت حل مسئله (پایه یازدهم) را مورد سنجش قرار داده است.

روش پژوهش

مشارکت‌کنندگان

با اســـتفاده از روش نمونه‌گیری طبقه‌ایِ خوشـــه‌ایِ دومرحله‌ای[1] نمونه معرفـــی برای تعیین عملکرد ریاضی دانش‌آموزان پایه ششم شهر تهران انتخاب شد. در مرحله اول نمونه‌گیری، مدرسه‌ها بر اساس طبقات تعیین‌شده، انتخاب شدند. در مرحله دوم و پس از تعیین هر یک از مدرسه‌های هدف، فهرست کلاس‌های ششم هر مدرسه، استخراج و از بین کلاس‌های ششم آن مدرسه یکی به‌طور تصادفی انتخاب شد و همه دانش‌آموزان آن کلاس به‌عنوان نمونه مطالعه انتخاب شـــدند. برای اجرای طبقه‌بندی مدرسه‌ها از دو نوع طبقه‌بندیِ صریح و ضمنی (لاروشه و همکاران[2]، ۲۰۱۶) استفاده شد. مناطق آموزشی به‌عنوان طبقه صریحِ نمونه‌گیری[3] به‌منظور تشکیل چارچوب‌های جداگانه برای تحلیل در نظر گرفته شـــد و دو متغیر جنسیت و نوع مدرسه به‌عنوان متغیرهای طبقه‌بندیِ ضمنی[4] به کار رفت که نقش اصلی آنها در مرتب کردن مدرســـه‌های درون هر یک از مناطق بود. این کار باعث شد هر یک از متغیرهای طبقه‌بندی نمونهٔ مدرسه‌ها به جامعه مشابهت بیشتری داشته باشد و درنتیجهٔ نمونه‌گیری دقیق‌تر شـــود. علاوه‌براین، برای انتخاب مدرسه‌ها در هر یک از مناطق، از روش تصادفیِ منظمِ احتمالات متناسب با حجم[5] استفاده شد. همچنین، برای رفع مشکلِ انتخاب با احتمالات نامساویِ نمونه از وزن‌های نمونه‌گیری[6] (روتوفسکی و همکاران[7]، ۲۰۱۰) استفاده شد. به‌کارگیری این روش باعث می‌شود که احتمال انتخاب دانش‌آموزان تحت تأثیر بزرگی یا کوچکی مدرسه‌ای که در آن تحصیل می‌کنند با بقیه متفاوت نباشد و همهٔ دانش‌آموزان احتمال برابری در انتخاب داشته باشند. ۹۷۲۰ دانش‌آموز پایه ششم از ۳۲۵ مدرسه

1. Stratified Two-Stage Cluster Sample Design
2. LaRoche, Joncas & Foy
3. Implicit stratification
4. Explicit stratification
5. probabilities proportional to their size (PPS)
6. sampling weights
7. Rutkowski, Gonzalez, Joncas & von Davier

در مطالعه مشارکت داشتند که ۴۷۳۴ نفر (۴۸/۷ درصد) از نمونه دختر و ۴۹۸۶ نفر (۵۱/۳ درصد) پسر بودند.

ابزار پژوهش

آزمون سنجش یادگیری دانش‌آموزان برای پایش کیفیت یادگیری دانش‌آموزان پایهٔ ششم شهر تهران طراحی شد. در طراحی آزمون و ساخت سؤال‌های ریاضی، ابتدا محدوده دانش و مهارت‌های مورد نیازِ سنجش ریاضی در قالب چارچوب سنجش مورد توافق قرار گرفت و سپس سؤال‌هایی برای هر مبحث طراحی شد که شامل ۲۳۴ سؤال بود. علاوه‌براین، ۵۳ سؤال ریاضی از سؤال‌های منتشر شده مطالعه تیمز در دو سال ۲۰۱۵ و ۲۰۱۱ به دو منظورِ پوشش سطوح پایین عملکرد دانش‌آموزان در پایه چهارم و همچنین پیوند دادن مقیاس نمره‌های این مطالعه به مقیاس مورد استفاده در مطالعه تیمز، انتخاب و به مجموع سؤال‌ها اضافه شد. این مرحله به‌عنوان یکی از تأکیدات روش نقطه‌گذاری معیار مورد توجه قرار گرفت. بدین ترتیب، ۲۸۷ سؤال به‌عنوان سؤال‌های اصلی مطالعه در نظر گرفته شد. از ۲۸۷ سؤال آزمون ریاضی، ۱۳۱ سؤال (۴۵/۶ درصد) را سؤال‌های چندگزینه‌ای و ۱۵۶ سؤال (۵۴/۳ درصد) را سؤال‌های بازپاسخ تشکیل می‌دادند. چنین حجمی از سؤال‌ها قابل پاسخ‌گویی توسط هیچ‌یک از دانش‌آموزان نبود. برای رفع مشکل، با استفاده از نمونه‌گیری ماتریسی[1]، سؤال‌ها در ۲۰ دفترچه جداگانه توزیع شد. در توزیع سؤال‌ها به متوازن بودن دفترچه‌ها از نظر حیطه‌های شناختی و محتوایی و هم‌چنین نوع سؤال‌ها توجه شد.

پس از اجرای آزمون، هر یک از سؤال‌ها در معرض تحلیل‌های متعددی قرار گرفت که کیفیت سؤال‌های نهایی را مشخص کند. شاخص‌های مورد بررسی از مجموعه تحلیل‌های کلاسیک آزمون‌سازی شامل ضریب تمیز، درصد پاسخ‌گویی به سؤال و ضریب دشواری، درصد پاسخگویی به هر یک از گزینه‌ها (سؤال‌های چندگزینه‌ای) یا کدهای نمره‌گذاری (سؤال‌های بازپاسخ)، همبستگی‌های دو رشته‌ای نقطه‌ای برای هر یک از گزینه‌ها و کدهای نمره‌گذاری و شاخص‌ها از مجموعه تحلیل‌های نظریه پرسش پاسخ، شامل برآورد جایگاه از مجموعه تحلیل‌های مدل راش و تعیین ضرایب جایگاه، شیب و حدس به همراه میزان خطا و آگاهی هر سؤال بودند. مجموع این تحلیل‌ها اشکالات برخی از سؤال‌ها را مشخص کرد که با تمهیداتی از قبیل ترکیب سؤال‌های چندبخشی، تبدیل سؤال‌های دو نمره‌ای به یک نمره‌ای و حذف سؤال‌های نامطلوب، تنها سؤال‌های با کیفیت مناسب برای تحلیل‌های نهایی درمجموع سؤال‌ها باقی ماندند. نتیجه این تحلیل، باقی ماندن ۲۶۴ سؤال و حذف ۲۴ سؤال از مجموع سؤال‌ها بود. درنهایت، اعتبار سؤال‌های هر دفترچه با روش آلفای کرونباخ نیز محاسبه شد. اعتبار دفترچه‌های ۱ تا ۲۰ به ترتیب عبارت بودند از: ۰/۸۵، ۰/۸۶، ۰/۶۱، ۰/۸۷، ۰/۷۵، ۰/۸۸، ۰/۸۹، ۰/۸۹، ۰/۸۶، ۰/۸۴، ۰/۷۶، ۰/۸۵، ۰/۸۹، ۰/۸۸، ۰/۹، ۰/۸۸، ۰/۹۰، ۰/۸۲، ۰/۸۹ و ۰/۸۵.

1. matrix sampling

روش تجزیه‌وتحلیل

برای اینکه نتایج این مطالعه قابلیت مقایسه بیشتری با مطالعات قبلی در زمینه ریاضی داشته باشد و یکی از مراحل روش نقطه‌گذاری معیار رعایت شود، از برآوردهای سؤال‌های مشترک بین این مطالعه و مطالعه تیمز برای ایجاد پیوندددهی بین این دو مطالعه استفاده شد. این کار باعث شد نتایج این مطالعه در همان مقیاس تیمز ۲۰۱۵ در پایه چهارم ارائه شـــود. این فرایند از طریق به‌کارگیری روش انتقال خطی[1] از مجموعه تکنیک‌های پیونددهی[2] و هم‌ترازســـازی آزمون (فوی و ین[3]، ۲۰۱۶) انجام گرفت. با توجه به اینکه مقیاس لوجیت[4] برای محاســـبه برآورد توانایی به کار می‌رود، به‌منظور پرهیز از ارائه نمره‌ها به‌صورت منفی یا اعشاری از مقیاسی با میانگین ۵۰۰ و انحراف معیار ۱۰۰ استفاده شد. در این مقیاس، نمره ۵۰۰ معادل با میانگین نمره‌های مقیاس در مطالعه و نمره‌های کمتر یا بیشتر از ۵۰۰ به معنای نمره‌های کمتر یا بیشتر از میانگین مقیاسی است که میزان کمتر یا بیشتر شدن آن را می‌توان با استناد به انحراف معیار ۱۰۰ قیاس کرد.

توصیف انتظارات در هر یک از سطوح عملکردی

در طراحی سؤال‌های ریاضی، سطوح عملکردی مشخصی برای هر سؤال در نظر گرفته شد. چهار سطح عملکردی در طراحی سؤال‌ها مورد توجه بود و هر سؤال ناظر به یکی از این سطوح طراحی شد. تنظیم معیارهای عملکردی به‌صورت تراکمی در نظر گرفته شده است؛ بدین‌صورت که دانش‌آموزانی که به سطوح بالاتر عملکردی می‌رسند، لاجرم می‌توانند انتظارات سطوح پایین‌تر عملکردی را نیز اجرا کنند و بر آنان تسلط یابند. در این مطالعه، چهار سطح عملکردی شامل سطوح پایین، متوسط، بالا و پیشرفته در نظر گرفته شد.

پایین‌ترین سطح ناظر به مفاهیم و مباحث پوشش داده شده در پایه چهارم ابتدایی بود. در حقیقت، این سطح بررسی می‌کند که دانش‌آموزان تا چه اندازه به معیارهای مورد انتظار در پایه چهارم رسیده‌اند و به همین دلیل در این معیار، بیشتر از سؤال‌های مطالعه تیمز برای پوشش به محتوای این پایه استفاده شد. توصیف انتظارات در این معیار بدین شرح است: دانش‌آموزان باید توانایی خواندن و نوشتن اعداد طبیعی و اعداد مرکب (زمان) را داشـــته باشند، اعداد طبیعی را با یکدیگر مقایسه و مرتب کنند. با خواص عملیات (چهار عمل اصلی) آشنا بوده و قادر به تخمین محاسبات باشند. با مفهوم کسر و اعداد اعشاری (با یک رقم اعشار) آشنا باشند. جمع و تفریق کسرهایی که مخرج یکی مضربی از مخرج دیگری است را به دست آورند. برای اندازه‌گیری طول و سطح واحد مناسب اندازه‌گیری انتخاب کنند و طول پاره‌خط، مساحت سطح و حجم شکل را با واحدهای داده شده به‌طور تقریبی اندازه‌گیری کنند. خطوط موازی با و عمود بر هم را تشـــخیص داده و رسم کنند. با ویژگی‌های چندضلعی‌ها آشنا باشـــند. ویژگی‌های مکعب، مکعب مستطیل، استوانه و مخروط را بدانند و گسترده آنها را

1. linear transformation
2. linking
3. Foy & Yin
4. logit metric

بشناسند. انواع زاویه را بشناسند و زاویه‌ها را با هم مقایسه کنند. محیط و مساحت مربع، مستطیل و مثلث را پیدا کنند. تقارن محوری و شکل‌های متقارن را تشخیص دهند. همچنین در حیطه نمایش داده‌ها باید اطلاعات مورد نیاز برای حل مسئله را از روی جدول و نمودار دریافت کنند. با مفهوم شانس و احتمال ساده آشنا باشند. مسائل ساده مرتبط با مفاهیم این سطح را حل کنند.

سطح عملکردی بعدی مورد توجه (معیار متوسط) مربوط به مفاهیمی است که در پایه پنجم از دانش‌آموزان انتظار می‌رود. از لحاظ توصیفی انتظار می‌رود که دانش‌آموزان باید بتوانند در حیطه اعداد، اعداد طبیعی را از میان تعدادی عدد شناسایی کرده، گسترده یک عدد را نوشته و برعکس با داشتن گسترده یک عدد، آن را با رقم بنویسند، زوج یا فرد بودن عدد را تشخیص دهند، تقریب عددی یک عدد را به کمک قطع کردن با تقریب مشخص به دست آورند. با مفهوم مضرب آشنا باشند و مضارب یک عدد را تشخیص داده و بنویسند. کسرهای بزرگ‌تر از واحد یا اعداد مخلوط را تشخیص داده و به یکدیگر تبدیل کنند. تبدیل اعداد اعشاری‌ها به کسر و برعکس را انجام دهند. ارزش مکانی رقم‌های با یک عدد اعشاری را درک کرده و مشخص کنند. ترتیب عملیات در عبارت شـــامل چهار عمل اصلی را بشناسند و آن را به‌درستی به کار بگیرند. جمع و تفریق اعداد مخلوط و اعشاری‌ها (با انتقال) را محاسبه کنند. ضرب یک عدد طبیعی در یک کسر یا یک عدد اعشاری را انجام دهند. مفاهیم نســـبت و درصد را درک کرده و دو یا چند نســـبت را با هم مقایسه کنند. رابطه دو دنباله را تشخیص دهند. در حیطه اندازه‌گیری و اشکال هندسی، دانش‌آموزان باید بتوانند واحدهای سطح و حجم را شناخته و آنها را به هم تبدیل کنند. ویژگی‌های قطر در چهارضلعی‌ها را درک کنند. مجموع زاویه‌های داخلی چندضلعی را به دست آورند و محیط دایره را محاسبه کنند. همچنین در حیطه نمایش داده‌ها شاگردان باید مفهوم میانگین را درک کرده و میانگین چند داده گسسته را پیدا کنند. نمودار خط شکسته را بشناسند و اطلاعات موجود را از روی آن دریافت کنند. مفهوم فراوانی نسبی و ارتباط آن با نمودار دایره‌ای را درک کنند و اطلاعات موجود در نمودار دایره‌ای را دریافت کنند.

ســـطح عملکردی سوم یا همان معیار بالا متناسب با انتظارات معمول از دانش‌آموز پایه ششم بر اساس کتاب درســـی ریاضی آنان بود. به‌طور مشـــخص از آنان انتظار می‌رود که در حیطه اعداد بتوانند یک عدد طبیعی با ویژگی‌های مشخص را پیدا کنند، بخش‌پذیری یک عدد بر عددی دیگر را تشخیص دهند، الگوی بخش‌پذیری بر ۲، ۳، ۵ و ۹ را بیان کنند، تقریب عددی یک عدد (عدد طبیعی) را به کمک گرد کردن با تقریب مشخص به دســـت آورند. مفهوم ضرب عدد در کسر، دو کسر و اعداد مخلوط در هم را درک کرده و انجام دهند. ضرب دو عدد اعشـــاری در هم را انجام دهند. تقســـیم یک کسر بر یک عدد طبیعی و برعکس را درک کرده و انجام دهند. معکوس یک عدد را درک کنند و از آن برای تقســـیم دو کسر بر هم استفاده کنند. مفهوم عدد صحیح را درک کننـــد و اعداد صحیح و قرینه آن را روی محور نمایش دهند. اعداد صحیح را مقایســـه کرده و مرتب

کنند. مفهوم تناسب را درک کنند. کمیت‌های متناسب را تشخیص دهند. جزء مجهول در تناسب را به دست آورند. تبدیل نسبت به درصد و برعکس را انجام دهند. جمله عمومی یک دنباله را به کمک عبارت‌های کلامی بنویسند. مجهول در تساوی‌ها و نامساوی‌های شامل عبارت‌های عددی را پیدا کنند. در حیطه اندازه‌گیری و شکل‌های هندسی فاصله نقطه از خط را درک و عمودمنصف یک پاره‌خط را رسم کنند. ویژگی عمودمنصف را درک کنند. مفهوم نیمساز یک زاویه را درک کنند. روابط دو زاویه با هم (متمم، مکمل، متقابل به رأس) را درک کنند. گسترده احجام را تشخیص داده و رسم کنند و از روی گسترده احجام، شکل سه‌بعدی را تشخیص دهند. مختصات نقطه در ربع اول را بشناسند و نقطه‌ای با مختصات داده شده را در ناحیه مشخص کنند. در حیطه نمایش داده‌ها، باید نمودار خط شکسته را رسم کنند. اطلاعات نمودارهای مختلف ارائه‌شده از یک وضعیت را با هم مقایسه کنند.

آخرین سطح عملکردی مورد انتظار از دانش‌آموزان معیار پیشرفته بود که به‌گونه‌ای طراحی شده بود که دانش‌آموزان پیشرفته پایه ششم را به‌خوبی اندازه‌گیری کند. به‌طور تفصیلی‌تر، از دانش‌آموزان در این معیار انتظار می‌رود که در حیطه اعداد، تقریب عددی یک عدد (عدد اعشاری) را به کمک گرد کردن با تقریب مشخص به دست آورند. تقسیم یک عدد اعشاری بر یک عدد طبیعی و برعکس و همچنین تقسیم دو عدد اعشاری با تقریب داده شده را انجام دهند. مجموعه‌ای از اعداد داده شده (صحیح، اعشاری، مخلوط) را مقایسه و مرتب کنند. با یافتن کوچک‌ترین مضرب مشترک دو یا چند عدد، کوچک‌ترین مخرج مشترک دو کسر را به دست آورند. در مسائل درصد (پیدا کردن درصدی از یک کل، پیدا کردن کل با داشتن درصد و پیدا کردن درصد با داشتن کل و جزء) مجهول را پیدا کنند. در حیطه اندازه‌گیری و شکل‌های هندسی مساحت دایره را پیدا کنند. تقارن چرخشی و دوران را درک کرده و شکل‌های متقارن را شناسایی و رسم کنند. نماهای سه‌گانه احجام را تشخیص دهند. مساحت کل و حجم مکعب و مکعب مستطیل را پیدا کنند. مختصات قرینه یک نقطه (یا شکل) را نسبت به محور تقارن (در ربع اول) پیدا کنند. یک نقطه (یا شکل) را با داشتن بردار انتقال (در ربع اول) انتقال داده و برعکس با داشتن شکل انتقال داده شده، بردار انتقال را پیدا کنند. در حیطه نمایش داده‌ها، شاگردان باید نمودار دایره‌ای را رسم کنند. با داشتن یک نمودار دیگری برای همان وضعیت رسم کنند.

یافته‌ها

اولین روش به کار برده شده در این مطالعه، روش نقطه‌گذاری معیار است. چهار سطح عملکردی به پیروی از مطالعه تیمز تعریف شدند که شامل سطح عملکردی پایین با نقطه برش ۴۰۰، متوسط با نقطه برش ۴۷۵، بالا با نقطه برش ۵۵۰ و پیشرفته با نقطه برش ۶۲۵ بودند. نمره‌های دانش‌آموزان از طریق میانگین پنج نمره مقادیر

محتملی محاسبه شـــد که برای هر دانش‌آموز برآورد شده بود. مقادیر محتمل[1] از تلفیق داده‌های پیشینه‌ای دانش‌آموزان (که عمدتاً از پرسشـــنامه‌های پیشـــینه‌ای دانش‌آموزان به دست می‌آید و با کدگذاری تصنعی به داده‌های دوبخشی تبدیل شده است) با داده‌های آزمون، پنج نمره پسین را برای هر دانش‌آموز به‌گونه‌ای تولید می‌کند که علاوه بر امکان بررسی خطاهای بیشتر، نزدیک‌ترین حالت را با برآوردهای جامعه داشته باشند. بر اســـاس این نقاط برش مشخص شد ۷۵ درصد از دانش‌آموزان به سطح عملکردی پایین رسیده‌اند، ۴۸ درصد از آنان به ســـطح متوسط دست یافته‌اند، ۱۸ درصد سطح بالا را کسب کرده‌اند و تنها ۲ درصد از دانش‌آموزان به ســـطح عملکردی پیشرفته رســـیده بودند. البته در توضیح توجه داشته باشید که درصدهای فوق به‌صورت تجمعی در نظر گرفته شده‌اند، به این معنا که دانش‌آموزانی که در سطوح بالایی قرار دارند، در سطوح پایین‌تر نیز محسوب شده‌اند.

برای اینکه کارکرد این روش مورد بررسی قرار گیرد، سطوح عملکردیِ منتسب شده به هر یک از ویژگی‌های مکعب، مکعب مســـتطیل، اســـتوانه و مخروط را بدانند و گستره آنها را بشناسند با روش نقطه‌گذاری معیار با سطح عملکردیِ اولیه‌ای که توسط کارشناسان موضوعی در نظر گرفته شده بود، مقایسه شد. بدین منظور، ابتدا آزمودنی‌های موجود در فاصله ۵ نمره از نقاط برش جدا شدند. تعداد آزمودنی‌های مرتبط با هر یک از سطوح عملکردی پایین ۳۴۹ نفر، سطح عملکردی متوسط ۴۳۶ نفر، سطح عملکردی بالا ۳۲۹ نفر و سطح عملکردی پیشرفته ۷۷ نفر بود. سپس میزان پاسخ‌گویی هر سؤال به تفکیک در هر یک از این چهار گروه مقایسه شد و بر این اساس، درجه عملکردیِ متناظر با هر سؤال بر اساس روش نقطه‌گذاری معیار مشخص شد. درنهایت، سطح انتسابیِ عملکردی سؤال با آنچه از قبل توسط کارشناسان موضوعی تعیین شده بود، مقایسه شد. نتایج در جدول (۱) ارائه شده است. در این جدول، دسته‌بندی ویژگی‌های مکعب، مکعب مستطیل، استوانه و مخروط را بدانند و گستره آنها را بشناسند در هر سطح عملکردی به دو بخش دقیق و تقریبی تفکیک شده است. سؤال‌هایی که به‌طور دقیق به سطح عملکردی اختصاص داده شده‌اند، واجد تمامی شروط هستند ولی سؤال‌هایی که به‌عنوان تقریبی طبقه‌بندی شده‌اند، برخی از شـــروط را ندارند (توضیحات بیشتر در بخش مقدمه مطرح شده است). به‌علاوه، برخی از سؤال‌هایی که نسبت اندکی از نمونه در سطح پیشرفته به آن پاسخ داده‌اند، به‌عنوان سؤال‌های خیلی دشوار طبقه‌بندی شده‌اند. در این جدول، درصدهای ارائه شده از کل سؤال‌ها محاسبه شده است.

1. plausible values

جدول (۱) مقایسه نتایج به‌دست‌آمده از روش نقطه‌گذاری معیار با دسته‌بندی اولیهٔ سؤال‌ها بر اساس نظر کارشناسان

دسته‌بندی بر اساس روش نقطه‌گذاری معیار	دسته‌بندی اولیه بر اساس نظر کارشناسان			
	پایین	متوسط	بالا	پیشرفته
دقیق	۱۱ (۴/۲٪)	۴(۱/۵٪)	۲ (۰/۸٪)	-
تقریبی	۷ (۲/۷٪)	۴(۱/۵٪)	۲ (۰/۸٪)	-
پایین	۱۸ (۶/۸٪)	۸ (۳/۰٪)	۴ (۱/۵٪)	۰ (۰/۰٪)
دقیق	۱۰(۳/۸٪)	۶ (۲/۳٪)	۳ (۱/۱٪)	۰ (۰/۰٪)
تقریبی	۱۳(۴/۹٪)	۸ (۳/۰٪)	۵ (۱/۹٪)	۱ (۰/۴٪)
متوسط	۲۳ (۸/۷٪)	۱۴ (۵/۳٪)	۸ (۳/۰٪)	۱ (۰/۴٪)
دقیق	۵(۱/۹٪)	۱۱ (۴/۲٪)	۶ (۲/۳٪)	۴ (۱/۵٪)
تقریبی	۸ (۳/۰٪)	۲۴ (۹/۱٪)	۶ (۲/۳٪)	۷ (۲/۷٪)
بالا	۱۳ (۴/۹٪)	۳۵ (۱۳/۳٪)	۱۲ (۴/۵٪)	۱۱ (۴/۲٪)
دقیق	۵(۱/۹٪)	۱۱ (۴/۲٪)	۲۱ (۸/۰٪)	۱۴ (۵/۳٪)
تقریبی	۴(۱/۵٪)	۱۳ (۴/۹٪)	۱۲ (۴/۵٪)	۵ (۱/۹٪)
پیشرفته	۹ (۳/۴٪)	۲۴ (۹/۱٪)	۳۳ (۱۲/۵٪)	۱۹ (۷/۲٪)
خیلی دشوار	-	۷ (۲/۷٪)	۹ (۳/۴٪)	۱۶ (۶/۱٪)
کل	۶۳ (۲۳/۸٪)	۸۸ (۳۳/۳٪)	۶۶ (۲۵/۰٪)	۴۷ (۱۷/۸٪)

جدول بالا نشان می‌دهد که ۶۳ سؤال (معادل ۲۳/۹ درصد از سؤال‌ها) در همان سطح عملکردی قرار گرفتند که توسط کارشناسان موضوعی تعیین شده بودند. ۳۲ سؤال (معادل ۱۲/۱ درصد از سؤال‌ها) به‌عنوان سؤال‌های خیلی دشوار شناخته شدند و بقیه سؤال‌ها در طبقاتی قرار گرفتند که با دسته‌بندی اولیهٔ آنان متفاوت بود. برای اینکه مشخص شود مابقی سؤال‌ها در سطوح عملکردی بالاتر یا پایین‌تر از سطح اولیه قرار گرفته‌اند، خلاصه نتایج ارائه شده در جدول (۲) مفید است. این جدول به تفکیکِ هر یک از سطوح عملکردی اولیه تنظیم شده و درصدهای مندرج در آن برای هر یک از سطوح محاسبه شده است.

جدول (۲) خلاصه چگونگی دقت روش نقطه‌گذاری معیار بر اساس دسته‌بندی اولیه کارشناسان

نتایج حاصل از روش نقطه‌گذاری معیار	دسته‌بندی اولیه بر اساس نظر کارشناسان			
	پایین	متوسط	بالا	پیشرفته
درست تشخیص داده	۱۱ (۱۷/۵٪)	۶ (۶/۸٪)	۶ (۹/۱٪)	۱۴ (۲۹/۸٪)
تقریباً درست تشخیص داده	۷ (۱۱/۱٪)	۸ (۹/۱٪)	۶ (۹/۱٪)	۵ (۱۰/۶٪)
برآورد بالاتر از سطح اولیه	۴۵ (۷۱/۴٪)	۵۹ (۶۷/۰٪)	۳۳ (۵۰/۰٪)	۰ (۰/۰٪)
برآورد پایین‌تر از سطح اولیه	۰ (۰/۰٪)	۸ (۹/۱٪)	۱۲ (۱۸/۲٪)	۱۲ (۲۵/۵٪)
سؤال‌های خیلی دشوار	۰ (۰/۰٪)	۷ (۸/۰٪)	۹ (۱۳/۶٪)	۱۶ (۳۴/۰٪)

در این جدول، مشخص است که در روش نقطه‌گذاری معیار، بیشتر سؤال‌ها با سطح اولیهٔ خود متفاوت بوده و بالاتر از آن بودند. سطوح اولیهٔ سؤال‌ها توسط کارشناسان موضوعی تعیین شده بود. علاوه‌براین، در دو سطح عملکردی پایین و پیشرفته بیشـــترین مطابقت بین روش نقطه‌گذاری معیار و بررسی محتوایی کارشناسان وجود دارد. در بخش دوم، ســـؤال‌ها بر اساس روش علامت‌گذاری تحلیل شده اســـت. در این روش به جای اجرای صرف علامت‌گذاری و مشخص کردن نمره‌های متناظر با هر یک از علامت‌ها، ابتدا سطح عملکردی هر سؤال بر اساس نظر کارشناسان تعیین شده و سپس آستانه هر سطح عملکردی بر اساس تجمیع سؤال‌ها و ترکیب آماره‌های آنها، به‌ویژه پارامتر جایگاه سؤال، تعیین می‌شود. این روش در نرم‌افزار IATA در پیش گرفته می‌شود و به نظر می‌رسد نسبت به روش سنتی علامت‌گذاری از نوسانات مربوط به سؤال‌های مرتب‌شده غیر هم‌سطح، فارغ خواهد بود.

جدول (۳) خلاصه آماره‌های پارامتر جایگاه سؤال‌های قرار گرفته در سطوح عملکردی

سطح عملکردی	میانگین پارامتر جایگاه	انحراف معیار پارامتر جایگاه
پایین	۰/۰۷	۰/۹۴
متوسط	۰/۸۳	۰/۷۸
بالا	۱/۳۵	۰/۷۴
پیشرفته	۱/۸۰	۰/۹۴

فاصله بین میانگین‌های متوالی از ۰/۴۵ تا ۰/۷۶ در نوسان است که کمتر از انحراف معیار پارامترهای جایگاه سؤال‌ها در هر یک از سطوح عملکردی است. از این مقایسه می‌توان متوجه شد که پراکندگی پارامتر سؤال‌های

درون هر یک از طبقات زیاد اســـت و مانع ایجاد طبقه‌بندی‌های دقیق شده است. البته به‌عنوان یکی از دلایل احتمالی می‌توان به سطح‌بندی اولیهٔ نامناسب سؤال‌ها توسط کارشناسان نیز اشاره کرد. به‌منظور مقایسه انتخاب احتمال پاسخ‌های متفاوت و تأثیر آنها بر طبقه‌بندی دانش‌آموزان در روش نقطه‌گذاری معیار، پنج احتمال پاسخِ ۰/۵۲، ۰/۵۷، ۰/۶۲، ۰/۶۷ و ۰/۷۵ در نظر گرفته شـــد و تأثیر این احتمال پاسخ‌ها بر آستانه‌های به‌دست‌آمده بررسی شد. نتایج در جدول (۴) مطرح شده است.

جدول (٤) تأثیر پنج احتمال پاسخ متفاوت بر آستانه‌ها و درصد دانش‌آموزان قرار گرفته در سطوح عملکردی

سطح عملکردی		احتمال پاسخ				
		۰/۵۲	۰/۵۷	۰/۶۲	۰/۶۷	۰/۷۵
پایین	آستانه	-۰/۰۸	۰/۱۰	۰/۲۸	۰/۴۷	۰/۷۴
	درصد دانش‌آموزان	**۴۰/۴**	۳۳/۲	۲۵/۴	۱۷/۹	۹/۵
متوسط	آستانه	۰/۷۵	۰/۹۲	۱/۰۸	۱/۲۵	۱/۵۴
	درصد دانش‌آموزان	**۹/۲**	۵/۷	۲/۸	۱/۳	۰/۲
بالا	آستانه	۱/۲۸	۱/۴۴	۱/۵۵	۱/۷۲	۲/۰۱
	درصد دانش‌آموزان	**۱/۱**	۰/۴	۰/۲	۰	۰
پیشرفته	آستانه	۱/۷۶	۱/۹۱	۲/۰۵	۲/۲۰	۲/۴۶
	درصد دانش‌آموزان	۰	۰	۰	۰	۰

*نتیجه نهایی درصد دانش‌آموزان در روش علامت‌گذاری پررنگ‌تر شده است

درصد دانش‌آموزان در هر پنج احتمال پاسخِ در نظر گرفته شده با درصدهای حاصل‌شده از روش نقطه‌گذاری معیار تفاوت جدی دارند. از بین این احتمال پاســـخ‌ها، درصد دانش‌آموزان در احتمال پاســـخ ۰/۵۲ به درصد دانش‌آموزان روش نقطه‌گذاری معیار نزدیک‌تر است ولی بااین‌حال هنوز تفاوت زیاد است؛ به‌طوری‌که در روش نقطه‌گذاری معیار ۷۵، ۴۸، ۱۸ و ۲ درصد از دانش‌آموزان به سطوح عملکردی پایین، متوسط، بالا و پیشرفته می‌رسند ولی در احتمال پاسخ ۰/۵۲ در روش علامت‌گذاری، برای همین سطوح عملکردی به ترتیب ۴۰، ۹، ۱ و ۰ درصد از دانش‌آموزان است. علاوه‌براین، در دو احتمال پاسخ ۰/۶۷ و ۰/۷۵ در روش علامت‌گذاری، صفر درصد از دانش‌آموزان به سطوح عملکردی بالا و پیشرفته رسیده بودند که این موضوع دقت طبقه‌بندی را برای احتمال پاسخ‌های مشخص‌شده، کاهش می‌دهد.

بحث و نتیجه‌گیری

این مطالعه با هدف بررسی تأثیر استفاده از دو روش معیارگزینیِ نقطه‌گذاری معیار و علامت‌گذاری بر داده‌های مطالعه کلان‌مقیاس برنامه رصد کیفیت آموزشی شهر تهران (برکات) طرح‌ریزی شده است. از بین روش‌های متعدد معیارگزینی، این دو روش توسعه‌یافته‌تر است و در مطالعات کلان‌مقیاس استفاده می‌شوند. در حال حاضر در هر مطالعه کلان‌مقیاس از یکی از این روش‌ها استفاده می‌شود. به همین دلیل، این پرسش به وجود می‌آید که کدام‌یک از این روش‌ها برای مطالعه برکات که برای سنجش درس ریاضی دانش‌آموزان پایه ششم شهر تهران طرح‌ریزی شده بود، بهتر عمل می‌کند. بدین منظور، هر دو روش اجرا شده و نتایج آن مورد بررسی قرار گرفت.

نتایج نشان داد که استفاده از روش نقطه‌گذاری معیار باعث می‌شود که به ترتیب ۷۵، ۴۸، ۱۸ و ۲ درصد از دانش‌آموزان حداقل نمره‌های لازم را در سطوح عملکردی پایین، متوسط، بالا و پیشرفته کسب کنند. علاوه‌براین، تبعیت از روش نقطه‌گذاری معیار و سطح‌بندی کردن مجدد سؤال‌ها با این روش نشان داد که ۲۳/۹ درصد از سؤال‌ها در همان سطحی قرار می‌گیرند که توسط کارشناسان موضوعی تعیین شده بودند. در مقابل، استفاده از روش علامت‌گذاری با استفاده از نرم‌افزار IATA فاصله ۰/۴۵ تا ۰/۷۶ بین میانگین‌های پارامتر جایگاه را نشان داد که مقایسه آنها با انحراف معیار پارامترهای جایگاه در هر سطح نشان می‌دهد که دسته‌بندی اولیهٔ کارشناسان موضوعی با داده‌ها تناسب زیادی ندارد. همچنین، تأثیر استفاده از پنج احتمال پاسخِ ۰/۵۲، ۰/۵۷، ۰/۶۲، ۰/۶۷ و ۰/۷۵ نشان داد که با وجود تأکید بر احتمال پاسخِ ۰/۶۷، کمترین احتمال پاسخ (۰/۵۲) نتایج واقعی‌تری را تولید می‌کند.

اگر بخواهیم یک ملاک خارجی را برای ارزیابی نتایج حاصل از معیارگزینی در نظر بگیریم، مطالعه تیمز ۲۰۱۹ که هم‌زمان با مطالعه برکات گردآوری شده است، ملاک مناسبی خواهد بود. بر اساس نتایج مطالعه تیمز ۲۰۱۹ در درس ریاضی به ترتیب ۶۸، ۳۹، ۱۳ و ۲ درصد از دانش‌آموزان پایه چهارم و ۶۸، ۳۷، ۱۴ و ۳ درصد از دانش‌آموزان پایه هشتم به ترتیب در سطوح عملکردی پایین، متوسط، بالا و پیشرفته قرار گرفته بودند (کبیری، زیر چاپ). با توجه به اینکه مطالعه برکات روی دانش‌آموزان پایه ششم شهر تهران اجرا شده است، نباید درصدهای دانش‌آموزان در سطوح عملکردی این مطالعه، تفاوت چشم‌گیری با مطالعه تیمز داشته باشد. از این بابت می‌توان نتیجه گرفت که برای معیارگزینی مطالعه برکات، روش نقطه‌گذاری معیار بهترین انتخاب است. بااین‌حال، همچنان این تردید وجود دارد که مشابهت نتایج روش نقطه‌گذاری معیار در این مطالعه با نتایج مطالعه تیمز به دلیل استفاده هر دو از یک روش یکسان بوده است.

نکته دیگر در مقایسه نتایج حاصل از روش نقطه‌گذاری معیار با روش علامت‌گذاری به چگونگی انتخاب

فاصله بین آستانه‌ها است. به نظر می‌رسد انتخاب فاصله‌های کاملاً مساوی در روش نقطه‌گذاری معیار کمی تصنعی باشد و در حالت‌های طبیعی، معیارها نباید چنین الگوهای یکنواختی داشته باشند. ازاین‌رو، درصورتی‌که روش علامت‌گذاری بتواند نتایج قابل دفاعی از لحاظ مشابهت با معیارهای بیرونی ارائه دهد، نسبت به روش نقطه‌گذاری معیار مرجح است.

با اینکه روش معیارگزینی اساساً یک روش ملاک مرجع است ولی توجه به هنجارها هم می‌تواند تا اندازه‌ای در انتخاب بهترین روش آن مؤثر واقع شود. چنانچه کارترایت (۲۰۱۵) بیان کرد نسبت کمِ افراد درون سطوح عملکردی باعث تولید آماره‌های بی‌ثبات و غیرقابل تفسیر خواهد شد. این توجه به‌ویژه در روش علامت‌گذاری که انتخاب احتمال‌های پاسخ بر معیارها و آستانه‌های سطوح عملکردی تأثیر می‌گذارد، اهمیت دارد. اینکه از بین مقادیر متفاوتِ احتمال پاسخ، دو مورد به درصد پاسخ‌گویی صفر در دو سطح عملکردی می‌انجامند، نشان می‌دهد که نمی‌توان به نتایج به‌دست‌آمده صرفاً نگاه ملاکی داشته و به هنجارها بی‌توجه بود.

بر اساس نتایج تجربی نمی‌توان برتری یکی از روش‌های معیارگزینی نسبت به دیگری را نشان داد (سیزک و بانچ، ۲۰۰۷)، بلکه برتری روش‌ها بیشتر به سهولت استفاده و مطابقت با شرایط اجرایی استفاده از آن است. روش علامت‌گذاری، مستلزم دقت بیشتر روی محتوای سؤال‌ها و بهره‌گیری از گروهی از متخصصان موضوعی است که درنتیجه هزینه‌های اجرا را افزایش داده و زمان بیشتری را می‌طلبد، درحالی‌که این محدودیت، کمتر در روش نقطه‌گذاری معیار گریبان‌گیر پژوهشگران است. علاوه‌براین، امکان توجیه این روش برای ذی‌نفعان و سیاست‌گذاران کم‌آشنا با جنبه‌های فنی آزمون‌سازی، روش نقطه‌گذاری معیار به دلیل استفاده از فواصل طبقاتیِ مساوی قابل درک‌تر است. از سوی دیگر، روش علامت‌گذاری نسبت به روش نقطه‌گذاری معیار با سؤال‌های آزمون مطابقت بیشتر دارد و به همین دلیل در بسیاری از سنجش‌های ملی به کار گرفته می‌شود (لوئیز و همکاران، ۲۰۱۲)؛ به‌طوری‌که در مواقعی مشاهده شده است که سؤال‌های زیادی در برخی از سطوح عملکردی در روش نقطه‌گذاری معیار دیده نمی‌شود (مولیس و همکاران، ۲۰۱۶). بنابراین، انتخاب بین این دو روش بیشتر به مقتضیات اجرایی و محیطی مربوط است.

درنهایت، به نظر می‌رسد باید به معیارگزینی به‌عنوان یکی از فنون کاربردی سنجش توجه بیشتری شود. این توجه، نه‌تنها در مطالعات کلان‌مقیاس بلکه در همه سنجش‌هایی لازم است که درجه‌بندی یا قبول و ردی یکی از تبعات شرکت در آزمون است. با شناختی که از فرایند آزمون‌سازی در داخل کشور موجود است، کمتر از معیارگزینی بهره برده می‌شود و معیارها نقاط از پیش تعریف شده‌ای هستند که حاصل مباحثات مدیریتی است تا محاسبات فنی و سنجشی. این مقاله و مقاله‌های نظیر آن کمک می‌کنند که

این روش به‌عنوان یک مبحث فنی، بیشتر در خدمت آزمون‌سازان و مدیران مراکز آزمون‌سازی قرار گیرد.

References

Cartwright, F. (2015). Item and test analysis. In G. Shiel & F. Cartwright (Eds.), *Analyzing data from a national assessment of educational achievement* (vol. 4) (pp. 125-257). Washington DC: International Bank for Reconstruction and Development / The World Bank.

Cizek, G., & Bunch, M. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. California: SAGE Publications, Inc.

Foy, P., & Yin, L. (2016). Scaling the TIMSS 2015 achievement data. In M. Martin, I. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS* 2015 (pp. 13.11-13.63). Boston: TIMSS & PIRLS International Study Center.

Habibzadeh, S., Delavar, A., Farrokhi, N., Minaei, A., & Jalili, M. (2019). The use of Rasch and item mapping in determining cut score of comprehensive pre internship exam. *Research in Medicine Education, 11*(3), 59-70. [in Persian].

Jalili, M., & Mortaz Hejri, S. (2012). Standard setting for objective structured clinical exam using four methods: Pre-fixed score, Angoff, borderline regression and Cohen. *Strides in Development of Medical Education, 9*(1), 77-84. [in Persian].

Jalalizadeh, M., Delavar, A., Farokhi, N., & Askari, M. (2020). Comparison of ANGOF-based IRT method and Bookmark method for standard Setting of MSRT language test. *Journal of Research in Teaching, 7*(4), 49-69. [in Persian]

Kabiri, M. (2019). BARAKAAT: Program for monitoring educational quality in Tehran – Iran's first provincial large-scale assessment, UNESCO-NEQMAP website. Retrieved in https://neqmap.bangkok.unesco.org/barakaat-program-for-monitoring-educational-quality-in-tehran-irans-first-provincial-large-scale-assessment/

Kabiri, M. (2020). *Program for monitoring educational quality in Tehran (BARAKAAT): specifying the quality of math education in 6th grade* (vol. 1), Research report: Tehran city department of Education. [in Persian].

Kabiri, M. (in press). *Quality of math and science education in Iran, comparing with others coantries: Results of TIMSS 2019*, Tehran: Madresseh Pub. [in Persian].

Makarem, A., Mahdavifard, H., & Gholami, H. (2017). Evaluation of passing scores in semiotics: An objective structured clinical examination for medical students of Mashhad University of Medical Sciences. *Strides in Development of Medical Education, 14*(1), 42-50. [in Persian].

Mortaz Hejri, S., Jalili, M., & Labaf, A. (2012). Setting standard threshold scores for an objective structured clinical examination using Angoff method and assessing the impact of reality checking and discussion on actual scores. *Iranian Journal of Medical Education. 11*(8), 885-894. [in Persian].

LaRoche, S., Joncas, M., & Foy, P. (2016). Sample design in TIMSS 2015. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (pp. 3.1-3.37). Boston: TIMSS & PIRLS International Study Center and International Association for the Evaluation of Educational Achievement (IEA).

Lewis, D. M., Mitzel, H. C., & Schulz, M. (2012). The bookmark standard setting procedure. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 225-254): Routledge.

Lissitz, R. W. (2013). Standard setting: Past, present, and perhaps future. In M. Simon, K. Ercikan, & a. M. Rousseau (Eds.), *Improving large-scale assessment in education: Theory, issues, and practice* (pp. 154-174). New York: Routledge.

Mullis, I. V. S., Cotter, K. E., Centurino, V. A. S., Fishbein, B. G., & Liu, J. (2016). Using scale anchoring to interpret the TIMSS 2015 achievement scales. In I. V. S. M. M. O. Martin, & M. Hooper (Ed.), *Methods and Procedures in TIMSS 2015* (pp. 14.11-14.47). Boston: TIMSS & PIRLS International Study Center and International Association for the Evaluation of Educational Achievement (IEA).

OECD. (2017). *PISA 2015 Technical Report*. Paris: OECD Publishing.

Olsen, R. V., & Nilsen, T. (2017). Standard setting in PISA and TIMSS and how these procedures can be used nationally. In S. Blömeke & J.-E. Gustafsson (Eds.), *Standard Setting in Education* (pp. 69-84): Springer.

Phillips, G. W. (2012). The benchmark method of standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 323-346): Routledge.

Price, L. R. (2017). *Psychometric methods: Theory into practice*. New York: Guilford Publications.

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data. *Educational Researcher, 39*(2), 142.

Shiel, G., & Cartwright, F. (2015). *National Assessments of Educational Achievement, Volume 4: Analyzing Data from a National Assessment of Educational Achievement*. Washington DC: The World Bank.

UIS. (2017). *Constructing UIS proficiency scales and linking to assessment to support SDG indicator 4.1.1 reporting*. Retrieved from UNESCO-UIS: http://uis.unesco.org/sites/default/files/documents/gaml4-constructing-uis-proficiency-scales-linking-assessments-support-sdg-indicator4.1.1-reporting.pdf