نمره‌دهی به آزمون‌های تشریحی: روش‌ها، چالش‌ها و راهکارها

اسحاق مرادی [1]، حسین دیده‌بان [2]*

تاریخ دریافت ۱۳۹۴/۰۵/۰۳ تاریخ پذیرش ۱۳۹۴/۰۷/۲۶

چکیده

پیش‌زمینه و هدف: تحقیقات نشان داده است که یادگیری دانشجویان تحت تأثیر مستقیم روش‌های ارزشیابی است. به اعتقاد بسیاری از پژوهشگران سؤال‌های تشریحی روش‌های برتر یادگیری را برمی‌انگیزند، اما دراین‌بین نحوه نمره‌دهی آزمون‌های تشریحی چالش بزرگی است که در بسیاری از موارد سبب بی‌اعتباری نمرات حاصل از آزمون می‌شود. برخلاف آزمون‌های عینی که اندازه‌گیری توانایی آزمون‌شوندگان مستقل ازنظر شخصی تصحیح کنندگان است در آزمون‌های تشریحی اندازه‌گیری توانایی موردسنجش به شخصی که پاسخ‌های سؤال‌ها را می‌خواند وابسته است بنابراین، دقت روش نمره‌دهی به آزمون‌های تشریحی به میزان زیادی به نحوه تصحیح اوراق و دقت عمل مصححان مربوط است. در این مقاله به بررسی شیوه‌های مختلف نمره‌دهی آزمون‌های تشریحی در مقالات مختلف می‌پردازیم و سعی داریم با بررسی عمیق متون، به روش استاندارد نمره‌دهی به آزمون‌های تشریحی بپردازیم.

مواد و روش‌ها: در این مطالعه مروری ابتدا پایگاه‌های مختلف داده‌ای Pubmed, Scopus,Proquest و مقالات منتشرشده در مجلات معتبر آموزش پزشکی شامل Academic medicine, British medical journal, teaching & learning in medicine, medical teacher, medical education و مجلات معتبر داخلی در زمینه آموزش پزشکی شامل مجله ایرانی آموزش در علوم پزشکی و گام‌های توسعه آموزش پزشکی با استفاده از کلیدواژه‌های آزمون تشریحی، نمره‌دهی، روش تحلیلی نمره‌دهی، روش کلی نمره‌دهی، روش ویژگی‌های اصلی و روش نمره‌دهی و معادل انگلیسی آن‌ها، مورد جستجو قرار گرفت و مقالات مرتبط انتخاب شدند.

یافته‌ها: در جستجوی منابع ابتدا تعداد ۲۴۰ مقاله و کتاب بدون محدودیت زمانی از جستجوی منابع اینترنتی به دست آمد که تعداد ۱۶۳ مطالعه به دلیل عدم ارتباط با موضوع از بررسی حذف گردید. تعداد ۲۳ مطالعه هم به دلیل وجود عنوان و نبود خلاصه از مطالعه خارج گردید. نهایتاً تعداد ۵۴ مطالعه جهت بررسی انتخاب گردید. بر اساس نتایج حاصل از جستجو، مطالب در چهار دسته زیر دسته‌بندی شدند:۱-تقسیم‌بندی‌های مختلف نمره‌دهی به سؤالات تشریحی ۲- انتخاب روش مناسب برای نمره‌دهی به سؤالات تشریحی ۳- چالش‌های نمره‌دهی به سؤالات تشریحی ۴-روش‌های جدید نمره‌دهی به سؤالات تشریحی

بحث و نتیجه‌گیری: ارتقای ارزیابی و نمره‌دهی به آزمون‌های تشریحی نیاز به استفاده از روش‌های مختلف و ترکیب روش‌های چندگانه و سطح بالا دارد. به‌هرحال اگر این اطمینان وجود دارد که استفاده از این آزمون‌ها، روش‌های برتر یادگیری در دانشجویان را ارتقا می‌دهد، هزینه‌های مربوط به نمره‌دهی و افزایش اعتبار نمرات آن را نیز باید بپردازیم.

کلیدواژه‌ها: آزمون‌های تشریحی، نمره‌دهی، ارزیابی دانشجو، چالش‌ها، راهکار

مجله دانشکده پرستاری و مامایی ارومیه، دوره سیزدهم، شماره هشتم، پی‌درپی ۷۳، آبان ۱۳۹۴، ص ۶۹۸-۶۹۲

آدرس مکاتبه: گروه آموزش پزشکی، دانشگاه علوم پزشکی تهران، تهران، ایران ، تلفن: ۰۹۳۵۴۰۱۲۴۱۴

Email: h-didehban@razi.tums.ac.ir

مقدمه

در دهه‌های اخیر استفاده از روش‌های ارزشیابی از دانشجویان پیشرفت‌های بسیاری داشته است (۱, ۲). دانشکده‌های پزشکی سعی در انجام ارزیابی‌های قابل‌اعتماد و پایا دارند که پزشکانی با مهارت‌های کافی به جامعه ارائه دهند (۳). بر همین اساس در فرایند آموزش، روش‌های مختلف ارزیابی از دانشجو مقوله‌ای است که ضرورتاً می‌بایست موردمطالعه قرار گیرد. تحقیقات زیادی این مهم را اذعان می‌کند که در بیشتر دانشکده‌های پزشکی استفاده از روش‌های کتبی

[1] عضو هیات علمی مرکز مطالعات و توسعه آموزش پزشکی، دانشگاه علوم پزشکی ارومیه، ارومیه، ایران

[2] دانشجوی دکترای آموزش پزشکی، گروه آموزش پزشکی، دانشگاه علوم پزشکی تهران، تهران، ایران (نویسنده مسئول)

ارزیابی امری رایج بوده و طیف گسترده‌ای از این آزمون‌ها در امر ارزیابی دانشجو مورداستفاده قرار می‌گیرد (۴). دراین‌بین، استفاده از روش آزمون‌های تشریحی که بیشترین تأکید را بر ساختن پاسخ از سوی یادگیرنده دارند با چالش‌های مختلفی از قبیل عدم نمونه‌گیری درست از آموخته‌های آزمون‌شوندگان و عدم اطمینان به نمره‌دهی از سوی مصححان روبرو هستند (۵-۹). چالش‌های مختلف موجود در امر ارزیابی دانشجو توسط آزمون‌های تشریحی، شیوه مناسب انتخاب روش نمره‌دهی و اثرات ناشی از روش نمره‌دهی، ضرورت پرداختن به این موضوع را به‌عنوان یکی از اولویت‌های حیطه ارزیابی، بیش‌ازپیش مورد تأکید قرار می‌دهد (۱۰-۱۲). اهمیت این چالش‌ها به حدی است که استفاده از فنّاوری‌های روز دنیا را برای نمره‌گذاری آزمون‌های تشریحی به بحث جدیدی دراین‌باره تبدیل کرده است(۱۳-۱۹) و به همین دلیل کارگاه‌های مختلفی برای بررسی روایی، پایایی و دیگر مشخصات آزمون‌ها در اغلب دانشگاه‌های معتبر پزشکی سراسر دنیا برگزار گردیده است. مقالات بسیاری به تشریح روش‌های نوین نمره‌دهی به آزمون‌های تشریحی پرداخته‌اند که استفاده از کامپیوتر در اولویت این روش‌ها قرار دارد (۲۰). مقالاتی به مزایا، معایب و مقایسه روش‌های مختلف نمره‌دهی این آزمون‌ها اشاره‌کرده‌اند (۸ و۲۱). و یک سری مقالات هم به بحث در رابطه با طراحی روش‌هایی برای نمره‌گذاری این آزمون‌ها پرداخته‌اند(۱۱). ولی ازآنجاکه نمره‌دهی به آزمون‌های تشریحی وابسته به توانایی شخص مصحح است دقت این نمرات از وابستگی زیادی به توانایی و مهارت شخص تصحیح گر برخوردار است و ازآنجاکه ارزیاب انسانی منبع خطای زیادی در نمره‌دهی است این مقاله ضمن تشریح چگونگی نمره‌دهی به سؤالات تشریحی به تشریح روش‌ها و رویکردهای جدید نمره‌دهی به این سؤالات می‌پردازد تا با دیدی وسیع سبب افزایش کیفیت نمرات داده‌شده به آزمون‌های تشریحی گردد.

مواد و روش کار

در این مطالعه مروری ابتدا پایگاه‌های داده شامل Google scholar,ERIC,Pubmed Scopus,Proquest و مقالات منتشرشده در مجلات معتبر آموزش پزشکی شامل Academic Medicine, British Medical Journal, Medical Teacher, Teaching & Learning In Medicine و Medical Education و مجلات معتبر داخلی در زمینه آموزش پزشکی شامل مجله ایرانی آموزش در علوم پزشکی و گام‌های توسعه آموزش پزشکی با استفاده از کلیدواژه‌های آزمون تشریحی[1] و نمره‌دهی[2]، روش تحلیلی نمره‌دهی[3]، روش کلی نمره‌دهی[4]، روش ویژگی‌های اصلی[5] و روش نمره‌دهی[6] مورد جستجو قرار گرفت و مقالات مرتبط انتخاب شدند.

یافته ها

ابتدا تعداد ۲۴۰ مقاله و کتاب بدون محدودیت زمانی از جستجوی منابع اینترنتی به دست آمد که تعداد ۱۶۳ مطالعه به دلیل عدم ارتباط با موضوع از بررسی حذف گردید. تعداد ۲۳ مطالعه هم به دلیل وجود عنوان و نبود خلاصه از مطالعه خارج گردید. نهایتاً تعداد ۵۴ مطالعه جهت بررسی انتخاب گردید. بر اساس نتایج حاصل از جستجو مطالب در چهار دسته زیر دسته‌بندی شدند (۸, ۱۶, ۲۱-۲۵):

۱- تقسیم‌بندی‌های مختلف نمره‌دهی به سؤالات تشریحی
۲- انتخاب روش مناسب برای نمره‌دهی به سؤالات تشریحی
۳- معایب و مزایای هر روش
۴- روش‌های جدید نمره‌دهی به سؤالات تشریحی.

۱- تقسیم‌بندی‌های مختلف نمره‌دهی به سؤالات تشریحی:

در سال‌های اخیر جهت‌گیری‌های جدیدی در زمینه روش‌های نمره‌دهی به سؤالات تشریحی مطرح و در حال گسترش است (۲۵, ۲۶). در این قسمت روش‌های مختلف نمره‌دهی به سؤالات تشریحی را مرور کرده و توضیح خواهیم داد.

الف-روش تحلیلی: در روش تحلیلی پاسخ فرد آزمون‌شونده به اجزای کوچک‌تری تقسیم و برای هر جز مشخص، نمره یا امتیاز جداگانه‌ای در نظر گرفته می‌شود(۲۱, ۲۷-۲۹).

ب- روش کلی: در این روش پاسخ نمونه به اجزا تقسیم نمی‌شود بلکه برداشت معلم از پاسخ به‌عنوان معیار اندازه‌گیری مورداستفاده قرار می‌گیرد. در این روش، معلم تمامی پاسخ هر فرد به یک سؤال را می‌خواند و یک برداشت کلی از آن کسب می‌کند و بعد این برداشت کلی را به یک نمره تبدیل می‌نماید. در این روش به هیچ عامل واحدی امتیاز خاصی داده نمی‌شود، بلکه همه عوامل موردنظر قرار می‌گیرند و کل پاسخ یک‌باره مورد قضاوت واقع می‌شود (۸, ۲۶, ۲۹).

ج-روش ویژگی‌های اصلی: در این روش مصحح ویژگی‌های اصلی پاسخ دانش‌آموز یا دانشجو به هر سؤال را می‌سنجد و برای

[1] Essay tests
[2] Scoring
[3] Analytical scoring
[4] Global scoring
[5] Primary traits
[6] Scoring rubric

هر یک از آن‌ها نمره‌های ۰ تا ۴ را که معرف عالی تا غیرقابل‌قبول‌اند منظور می‌نماید (۲۹). البته در بعضی از منابع این روش نمره‌دهی را یکی از روش‌های نمره‌دهی تحلیلی دانسته‌اند(۳۰).

۲-انتخاب روش مناسب برای نمره‌دهی به سؤالات تشریحی:

مقالات متعددی بهترین روش برای کاهش سوگیری و افزایش اعتبار نمرات آزمون‌های تشریحی را آموزش ارزیابان می‌دانند (۳۱). در مقایسه با روش‌های ذهنی نمره‌دهی، مقالات متعددی استفاده از مقیاس نمره‌دهی و روش تحلیلی را با توجه به معیارهای مشخص نمره‌دهی را در جهت افزایش اعتبار نمرات کاربردی‌تر می‌دانند(۲۱, ۲۳, ۳۲). در بعضی از مقالات نیز استفاده از روش نمره‌دهی تحلیلی به سؤالات را به دلیل اینکه دلایل توجیهی روشنی برای یادگیرنده ارائه می‌دهد، از روش کلی نمره‌دهی بهتر دانسته‌اند (۸, ۲۴, ۲۹) و از طرفی استفاده از روش کلی نمره‌دهی را آسان‌تر و سریع‌تر از روش نمره‌دهی تحلیلی دانسته‌اند (۸, ۳۳). عده‌ای از محققان نیز استفاده از هر دو روش و ترکیب دو روش کلی و تحلیلی را بهترین روش نمره‌دهی به سؤالات تشریحی دانسته‌اند. بدین‌صورت ابتدا سؤالات را به‌صورت کلی تصحیح و سپس پاسخ‌ها را به‌صورت تحلیلی تصحیح نمایند (۱۱, ۲۶). این روش نمره‌گذاری سبب جلوگیری از معایب هر دو روش نمره‌گذاری خواهد شد و اعتبار نمرات نیز بالا خواهد رفت (۲۹, ۳۰). بعضی از صاحب‌نظران نیز استفاده از روش نمره‌دهی کلی را زمانی که ارزیابی در مقیاس بزرگ و به‌صورت تراکمی انجام می‌شود توصیه می‌کنند و اشاره دارند که استفاده از این روش در مقایسه با روش تحلیلی تصحیح نمرات از هزینه پایینی برخوردار است (۳۳, ۳۴).

۳-چالش‌های نمره‌دهی به سؤالات تشریحی:

در مقالات به چالش‌های متفاوتی در رابطه با نمره‌دهی به سؤالات آزمون‌های تشریحی اشاره شده است. برخی از این چالش‌ها را می‌توان به موضوع اعتبار و پایایی نمرات (۳۳) و برخی دیگر نبود یک مقیاس استاندارد در روش نمره‌دهی کلی را از چالش‌های این روش نمره‌دهی دانسته‌اند و این‌گونه استدلال می‌کنند که این روش نمره‌دهی بیش از این‌که به نقاط ضعف آزمون‌شوندگان اشاره داشته باشد، به تمرکز نویسنده در مورد سؤال مربوطه اشاره دارد و از دادن بازخورد به فراگیرندگان در مورد اشتباهات رایج خود در مورد سؤال موردنظر عاجز است (۳۵). اما در مورد نمره‌دهی به روش تحلیلی چالشی که در مقالات به آن اشاره شده است، نبود معیار استاندارد نمره‌دهی در هر آزمون (۳۶)، ارزیابی جنبه‌های کیفی پاسخ به سؤالات ازجمله شیوه بیان و طرز استفاده از واژگان و در نظر نگرفتن معیارهای مهم دیگر پاسخ به سؤالات (۳۷) اشاره شده است. بعضی از مقالات نیز تنوع معیار و مقیاس نمره‌دهی و چگونگی استفاده ارزیاب از این مقیاس را سبب ایجاد نمرات متفاوت بین ارزیاب‌ها می‌دانند و به این نکته اشاره دارند که در یک برنامه آموزش ارزیابی و نمره‌دهی، افراد باید در جهت اهداف ارزیابی و استفاده از معیارهای مشابه و تلاش در جهت رتبه‌بندی کردن ارزیاب‌ها از تازه‌کار به متخصص قدم بردارند.(۳۸)

۴-روش‌های جدید نمره‌دهی به سؤالات تشریحی:

روش‌های جدید نمره‌دهی آزمون‌های تشریحی بیشتر به سمت کاهش سوگیری و افزایش اعتبار نمرات این آزمون با استفاده از روش‌های مبتنی بر کامپیوتر بوده است. شاید بتوان ایده اصلی این طرح را متعلق به آلیس پاجس در سال ۱۹۶۶ دانست. او این فرضیه را که می‌توان از کامپیوتر در نمره‌دهی به آزمون‌های تشریحی استفاده نمود را آغاز کرد ولیکن به دلیل عدم مقرون‌به‌صرفه بودن در آن سال‌ها این ایده مورد استقبال قرار نگرفت (۳۹). در سال‌های دهه ۱۹۹۰ با ظهور کامپیوترهای رومیزی این فرضیه دنبال شد و اوج این شکوفایی در سال ۲۰۱۲ با حمایت مالی بنیاد ملی هیولت۱ انجام شد این بنیاد در این سال با این فرضیه که سیستم نمره‌دهی مبتنی بر کامپیوتر آن‌ها توانایی ارائه نمرات قابل‌اعتماد و پایایی نسبت به ارزیاب انسانی ارائه می‌دهد تعداد هزار برگه امتحان تشریحی را به هشت ارزیاب انسانی ارائه داد و در آخر این فرضیه تائید گردید (۴۰). سیستم ارزیاب کامپیوتری مبتنی بر مدل‌های ریاضی بوده و برنامه‌های متعددی برای این سیستم مورداستفاده قرار گرفته است ولیکن پایه اصلی تمامی مدل‌های آن‌ها، اندازه‌گیری توافق بین ارزیاب درصورتی‌که نمرات بین دو ارزیاب با هم همخوانی نداشته باشد می‌باشد(۴۱). مقالات بسیاری در زمینه استفاده از سیستم‌های خودکار نمره‌دهی به آزمون‌های تشریحی منتشرشده است. محور این مقالات عموماً در زمینه اعتبار و پایایی روش خودکار نمره دهی(۵۲) و روش‌ها و مدل‌های بکار رفته در این روش(۲)، و مقایسه نمرات کسب‌شده از این روش با ارزیاب انسانی(۱۳،۵۳) می‌باشد. در تعدادی از مقالات نیز به چگونگی پیاده‌سازی این سیستم‌ها در نظام ارزشیابی خود اشاره‌کرده‌اند. به‌طورکلی تصحیح و نمره‌دهی از طریق این سیستم شامل ۴ مرحله: ۱-آماده‌سازی داده‌ها: در این مرحله نسخه الکترونیکی از پاسخ آزمون دهنده ایجاد می‌گردد.۲-استخراج ویژگی‌ها: در این مرحله فهرستی از ویژگی‌های که می‌تواند توسط کامپیوتر برای توصیف داده‌های ورودی مورداستفاده قرار گیرد استخراج می‌گردد.۳-فراگیری

[1] The William and Flora Hewlett Foundation

ماشین: در این مرحله با توجه به مشخصات مرحله دوم خروجی خاصی تولید می‌گردد. خروجی مورداستفاده برای اکثر برنامه‌های کاربردی نمره‌دهی خودکار شامل نمرات داده‌شده به پاسخ‌های تولیدشده توسط ارزیاب انسانی است.۴-طبقه‌بندی نمرات و تجزیه‌وتحلیل خطا: در این مرحله نمرات دسته‌بندی و تحلیل خطاها انجام می‌گیرد(۵۴).

بحث و نتیجه‌گیری

هرچند چالش‌های اخیر در زمینه ارزیابی دانشجویان پزشکی با آزمون‌های کتبی بیشتر به سمت افزایش عینیت نمره‌دهی آزمون‌های تشریحی بوده است ولی روند ارتقای روش‌های نمره‌دهی به این آزمون‌ها در مؤسسات آموزش پزشکی ایران، آهسته تراز دیگر نقاط دنیاست به‌گونه‌ای که در جستجوی منابع، مقالات اندکی به این موضوع پرداخته‌اند و بیشتر بر نتایج به‌کارگیری این نوع آزمون‌ها بر یادگیری و فرآیندهای آموزشی متمرکز بوده‌اند (۴۲)، درحالی‌که پیشرفت‌های زیادی دراین‌باره در دنیا رخ داده است(۴۳). بنابراین چالش‌های موجود در نمره‌دهی به این روش ارزیابی با توجه به استفاده روزافزون آن به‌اندازه فهم دیگر جزئیات هر یک از روش‌های ارزیابی دیگر از اهمیت خاصی برخوردار است. به‌طورکلی لازم است بر چگونگی، ساختار و نحوه ارتقا نمره‌دهی به آزمون‌های تشریحی تمرکز نمود تا تعیین گردد که آیا آزمون پیامدهای موردنظر یادگیری را ارزیابی می‌کند یا خیر (۴۴). لازم است از رویکردهای نمره‌دهی مبتنی بر کامپیوتر، برای نمره‌دهی به سؤالات تشریحی استفاده نمود تا قضاوت ارزشیابان به کمترین مقدار برسد(۲۲, ۴۵).

بااینکه در این مقاله به چالش‌هایی که بیشتر ارزیابان در قبال سؤالات تشریحی با آن‌ها روبرو هستند اشاره شد ولی سؤالات فراوانی در مورد ارزیابی توانمندی ارزیابان بی‌پاسخ مانده است (۴۶). در دهه‌های اخیر این حرکت با افزایش تعداد ارزیابان در آزمون‌های تشریحی (۴۷)، و استفاده از فنّاوری‌های جدید نمره‌دهی شروع‌شده تا بر روش‌های ارزیابی ذهنی و بدون اعتبار، کمتر تکیه شود (۴۸, ۴۹). ارزیابان، اعضای هیئت‌علمی و افراد دخیل باید درباره ضعف‌های ذاتی نمره‌دهی‌های سنتی به آزمون‌های تشریحی آگاه‌تر گردند. درعین‌حال باید توجه داشت که توانمندی ارزیابان در آزمون‌های تشریحی دارای ساختاری در هم تنیده است (۵۰) و برای افزایش اعتبار نمرات آزمون‌های تشریحی، به روش‌های چندگانه، ترکیبی و سطح بالا نیاز است (۸). در پایان، می‌توان گفت تصمیم‌گیری درباره چگونگی استفاده از روش‌های مختلف نمره‌دهی به سؤالات تشریحی، شرایطی که این آزمون‌ها در آن انجام می‌گیرد و نیز اینکه معیارها و استانداردهای نمره‌دهی به سؤالات تشریحی چگونه باشند (۵۱)، هنوز به‌عنوان چالش‌های مهمی در ارزیابی سؤالات تشریحی مطرح هستند.

References:

1. Griffin P. Assessment for Teaching. Cambridge University Press; 2014.
2. Miller AH, Imrie BW, Cox K. Student assessment in higher education: a handbook for assessing performance. Psychology Press; 1998.
3. Rheingold A, Seaman J, Berger R. Assessment across boundaries: How high-quality student work demonstrates achievement, shapes practice, and improves communities. Assessing Schools for Generation R (Responsibility) Springer; 2014. p. 115–31.
4. Nelson R, Dawson P. A contribution to the history of assessment: how a conversation simulator redeems Socratic Method. Assess Eval High Educ 2014;39(2):195-204.
5. Jonsson A, Svingby G. The use of scoring rubrics: Reliability, validity and educational consequences. Educ Res Rev 2007;2(2):130-44.
6. Johnson RL, Penny J, Gordon B, Shumate SR, Fisher SP. Resolving Score Differences in the Rating of Writing Samples: Does Discussion Improve the Accuracy of Scores? Language Assess Quart 2005;2(2):117-46.
7. Powers DE, Burstein JC, Chodorow M, Fowles ME, Kukich K. Stumping e-rater: Challenging the validity of automated essay scoring. Computers Human Behav 2002;18(2):103-34.
8. Bacha N. Writing evaluation: What can analytic versus holistic essay scoring tell us? System 2001;29(3):371-83.
9. Fullerton JT, Greener DL, Gross LJ. Scoring and setting pass/fail standards for an essay

certification examination in nurse-midwifery. Midwifery 1992;8(1):31-9.

10. Rudner L, Gagne P. An Overview of Three Approaches to Scoring Written Essays by Computer. Pract Assess Res eval 2001;7(26).
11. Coker DR. Improving Essay Tests: Structuring the Items and Scoring Responses. Clear Hou 1988;61(6):253-5.
12. Hughes DC, Keeling B. The Use of Model Essays to Reduce Context Effects in Essay Scoring. J Educ Measur 1984;21(3):277-81.
13. 1. Shermis MD. State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. Assessing Writing 2014;20:53–76.
14. Weigle SC. English language learners and automated scoring of essays: Critical considerations. Assessing Writing 2013;18(1):85-99.
15. Wang J. A new automated essay scoring: Teaching Resource Program; 2013. p. 608-11.
16. Ramineni C, Williamson DM. Automated essay scoring: Psychometric guidelines and practices. Assessing Writing 2013;18(1):25-39.
17. Dos Santos Figueira A, Gomes da Silva A, Machado de Melo B, Del Pino Lino A, Lobato L, Rossy F, et al. Module of evaluation automatic essay questions on virtual learning environment LabSQL. Information Systems and Technologies (CISTI), 2013 8th Iberian Conference on IEEE; 2013. p. 1–5.
18. Chen H, Xu J, He B. Automated Essay Scoring by Capturing Relative Writing Quality. Computer J 2013:bxt117.
19. Almond RG. Using Automated Essay Scores as an Anchor When Equating Constructed Response Writing Tests. Int J Test 2013;14(1):73-91.
20. Attali Y, Lewis W, Steier M. Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. Language Testing 2013;30(1):125-41.
21. East M. Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. Assessing Writing 2009;14(2):88-115.
22. Bridgeman B, Trapani C, Attali Y. Comparison of Human and Machine Scoring of Essays: Differences by Gender, Ethnicity, and Country. Appl Measur Educ 2012;25(1):27-40.
23. Adnan KAN. Effects of using a scoring guide on essay scores: Generalizability theory. Perceptual Motor Skills 2007;105(3 I):891-905.
24. Sachs SG, Reiser RA. An objective approach to scoring essays. J Instructional Devel 1979;3(4):19-22.
25. Harsch C, Martin G. Comparing holistic and analytic scoring methods: issues of validity and reliability. Assess Educ Princip Policy Practice 2012;20(3):281-307.
26. Nakamura Y. A comparison of holistic and analytic scoring methods in the assessment of writing. The Interface Between Interlanguage, Pragmatics and Assessment: Proceedings of the 3rd Annual JALT Pan-SIG Conference; 2004.
27. Espin CA, Weissenburger JW, Benson BJ. Assessing the writing performance of students in special education. Exceptionality 2004;12(1):55-66.
28. Chase CI. Essay test scoring: Interaction of relevant variables. J Educ Measur 1986;23(1):33-41.
29. Saif AA. Educational Measurement,Assessment and Evaluation. 3rd ed. Dowran; 2005.
30. Cooper CR, Odell L. Evaluating Writing: Describing, Measuring, Judging; 1977.
31. Sweedler-Brown CO. The effect of training on the appearance bias of holistic essay graders. J Res Devel Educ 1992;26(1):24-9.
32. Moskal BM, Leydens JA. Scoring rubric development: Validity and reliability. Practic Assess Res Eval 2000;7(10):1-11.
33. Huot B. Reliability, validity, and holistic scoring: What we know and what we need to know.

College Composition and Communication; 1990.P.201-13.

34. Myers M. A Procedure for Writing Assessment and Holistic Scoring. ERIC; 1980.
35. Elbow P. Ranking, Evaluating, Liking: Sorting Out Three Forms of Judgment. College English 1994;55(2):187-206.
36. Gamaroff R. Rater reliability in language assessment: the bug of all bears. System 2000;28(1):31-53.
37. Mertler CA. Designing scoring rubrics for your classroom. Practic Assess Res Evalu 2001;7(25):1-10.
38. Moskal BM. Scoring Rubrics: How?: ERIC Clearinghouse on Assessment and Evaluation. University of Maryland; 2000.
39. Shermis MD, Burstein J, Higgins D, Zechner K. Automated essay scoring. Writing assessment and instruction; 2010. p. 20-6.
40. Shermis MD, Burstein J. Handbook of automated essay evaluation: Current applications and new directions. Routledge; 2013.
41. Ben-Simon A, Bennett RE. Toward more substantively meaningful automated essay scoring. J Technol Learn Assess 2007;6(1).
42. Shakurnia A, Alijani H, Najjar S, komeili h, Elhampour H. The Effect of Two Assessment Methods on Exam Preparation and Study Strategies: Multiple Choice and Essay Questions. Iran J Med Educ 2013;13(4):306-18.
43. Zhang M, Williamson DM, Breyer FJ, Trapani C. Comparison of e-rater® Automated Essay Scoring Model Calibration Methods Based on Distributional Targets. Int J Test 2012;12(4):345-64.
44. Madu B, Ikeh EF. Effect of Scoring Patterns on Scorer Reliability in Economics Essay Tests. J Economics Sustainable Develop 2013;4(15):68-74.
45. Powers DE, Burstein JC, Chodorow MS, Fowles ME, Kukich K. Comparing the validity of automated and human scoring of essays. J Educ Comput Res 2002;26(4):407-25.
46. Myford CM, Wolfe EW. Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. J Educ Measur 2009;46(4):371-89.
47. Leckie G, Baird J-A. Rater Effects on Essay Scoring: A Multilevel Analysis of Severity Drift, Central Tendency, and Rater Experience. J Educ Measur 2011;48(4):399-418.
48. Attali Y. Method of model scaling for an automated essay scoring system. Google Patents; 2014.
49. Attali Y, Burstein J. Automated essay scoring with e-rater® V. 2. J Tech Learn Assess 2006;4(3).
50. Leckie G, Baird JA. Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. J Educ Measur 2011;48(4):399-418.
51. Ahn JY, Han KS, Mun GS. Evaluation of essay questions and applications for effective feedback. ICIC Express Letters 2013;7(3 A):793-8.
52. Yao X. Automated essay scoring. A comparative study; 2013. p. 650-3.
53. Hang M, Williamson DM, Breyer FJ, Trapani C. Comparison of e-rater® Automated Essay Scoring Model Calibration Methods Based on Distributional Targets. Int J Test 2012;12(4):345-64.
54. Gierl MJ, Latifi S, Lai H, Boulais A-P, Champlain A. Automated essay scoring and the future of educational assessment in medical education. Med Educ 2014;48(10):950–62.

Review Article

SCORING IN THE ESSAY TESTS QUESTIONS: METHODS, CHALLENGES AND STRATEGIES

Moradi E [1], Didehban H [2]

Received: 25 Jul, 2015; Accepted: 18 Oct, 2015

Abstract

Background & Aims: The related studies has shown that students learning is under the direct influence of assessment and evaluation methods. Many researchers believe that essay tests can assess the quality of the students' learning, however essay tests scoring a big challenge which causes their unreliability in many cases. Unlike objective tests that measure the examinees' ability independent of the individual judgment, in essay tests, the examinees' ability is directly related to the examiners' judgment. Therefore, the accuracy of test scoring is greatly related to the methods of paper correction, and the examiners accuracy. This study tries to explain various methods of essay tests scoring in different published papers.

Materials & Methods: In this study, first the different papers and articles published in national and international journals were selected by using keywords of essay test, scoring, and student assessment. Later, they were studied and analyzed.

Results*:* 240 articles without time limit were found by searching in the Web. 163 articles were excluded because of lack of relevant to the topic, and 23 articles were excluded because of lacking abstract. Finally, 54 articles were selected to be included in the study. Based on the search results of study, the search topic was grouped in four categories:1.defining different grading methods in essay questions 2.selecting appropriate methods for essay questions scoring 3.the challenges of essay questions scoring. 4. new methods of essay questions scoring.

Conclusion*:* Improving assessment and essay questions scoring requires using of various methods, and the combination of multiple high level tests. In applying these kinds of tests, we should pay attention to the reliability and validity of the essay tests and scores.

Keywords: Essay test, Scoring, Student assessment

Address: Department of Medical Education, Tehran University of Medical Sciences, Tehran, Iran
Tel: (+98) 09354012414
Email: h-didehban@razi.tums.ac.ir

[1] *Instructor, Education Development Center (EDC), Urmia University of Medical Sciences, Urmia, Iran*
[2] *MSc, PhD candidate in medical education, Department of Medical Education, Tehran University of Medical Sciences, Tehran, Iran (Corresponding Author)*