Outcomes in CME/CPD – Special Collection: Effect Size Benchmarking for Internet-based Enduring CME Activities

Jason J Olivieri and Mary Catherine Downes

Med-IQ, LLC, Baltimore, Maryland

ABSTRACT
The volume of certified, internet enduring materials produced per year has nearly doubled in the last decade. Meta-analyses indicate that Internet-based education for clinicians is effective; however, the relevance of these studies to the nearly 50,000 such activities certified per year is questionable. Effect size is one metric by which CME providers may assess effectiveness, but caution must be used in comparing effect size data with external benchmarks such as peer-reviewed literature. This report presents a pooled standardised mean difference (Cohen’s d) for 40 accredited, Internet-based enduring materials produced between 2016 and 2018. Data suggests that a Cohen’s d between 0.48 and 0.75 may be a useful benchmark. Benchmarks reported in the literature for this format are notably higher. The limitations of comparison to such benchmarks are considered.

Introduction
Are we doing a good job with internet-based continuing medical education (CME)? The number of certified, internet enduring materials produced per year has nearly doubled between 2010 and 2018[1]. Although growth suggests progress, meta-analysis has further indicated that internet-based CME can be effective in enhancing clinician knowledge, behaviour, and even patient health in comparison to no intervention[2]. In addition, Casebeer et al. studied 114 internet CME activities and reported a positive effect on the evidence-based decision-making of physicians[3]. Whether and how CME providers, responsible for nearly 50,000 internet-based enduring activities in 2018 alone, can compare their effectiveness to such published literature, as well as establish their own benchmarks, is the focus of this report.

One metric by which CME providers can compare the effectiveness of their CME to that of published literature is effect size. Although there are many different effect sizes, each represents a standardised measure of the magnitude of an intervention’s effect and, correspondingly, allows for the comparison of effect across similar interventions. The effect size measure most commonly employed in CME is Cohen’s d. The method for calculating has been described in a previous publication[4]. Cohen’s d is a relative measure, which means it requires a comparison to give it context. Unlike absolute measures, such as age or weight, which have meanings that are universally understood, a Cohen’s d value can only be deciphered in reference to another Cohen’s d. For initial reference, Cohen [5] recommended the following benchmarks: small effect (0.2), medium effect (0.5), and large effect (0.8). However, these benchmarks were intended to be temporary placeholders until more robust benchmarks could be established via repeated measurement in a given area. To that end, the aforementioned meta-analysis reported a pooled effect size (based on 126 studies) for internet-based education on clinician knowledge outcomes of 1.0[2] and Casebeer et al. [3] reported a pooled effect size of 0.82 across 114 internet-based enduring materials. Such benchmarks help orient our understanding of educational effect; however, direct comparisons are subject to numerous limitations, such as heterogeneity of clinician participants, educational techniques employed within the internet enduring format, assessment methods, and outcome measures, as well as the effect of publication bias.

The purpose of this study was to determine benchmarks of educational effectiveness for internet-based enduring materials relevant to CME providers. That is, effect size data for 40 activities produced by an accredited medical education company over a 3-year span were compared with preliminary benchmarks [5] and what has been reported in peer-reviewed literature [2,3].
Materials and Methods

Forty internet enduring materials launched between 6 May 2016, and 10 October 2018, were included in this study. An internet enduring material has been defined by the Accreditation Council for Continuing Medical Education as follows:

An internet enduring material is an online enduring activity that can be accessed whenever the learner chooses to complete it. The content can be accessed at any point during the lifespan of the activity and there is no specific time designated for participation. Examples include online interactive educational modules, recorded presentations, podcasts. [6]

Under this umbrella, the internet-based enduring materials included in this study fell into two categories: 1) digital publication (n = 31) or 2) webcast (n = 9). Digital publications consisted of text-based presentation of clinical data; supporting figures, graphs, tables, and references; and embedded polling questions. Webcasts consisted of slide-based presentations of clinical data and demonstration videos, as well as polling questions. Each activity in this study sample was certified for American Medical Association Physician’s Recognition Award (AMA PRA) Category 1 Credit™ and endured for 1 year.

Activity content spread across 13 distinct content areas. The most common were infectious disease (38%, n = 15), oncology (13%, n = 5), endocrinology (8%, n = 3), and cardiology (8%, n = 3). The remaining content areas included addiction medicine (n = 1), dermatology (n = 2), gastroenterology (n = 2), genetics (n = 2), immunology (n = 2), musculoskeletal disorders (n = 1), neurology (n = 1), psychiatry (n = 1), and pulmonology (n = 2). Thirty-five per cent of activities (n = 14) targeted solely primary care practitioners, and the remainder were developed for specialists (n = 20) or a mixed specialty-primary care audience (n = 6).

A one-group, pre- versus post-activity design was used to measure the educational impact of each activity. Pre- and post-activity assessment consisted of identical multiple-choice, single correct answer questions. All questions were developed by PhD- or master’s-level medical writers in collaboration with activity faculty and in accordance with National Board of Medical Examiners guidelines[7]. The median number of pre- and post-assessments per activity was 4 (range: 4–6). A per cent correct score (pre- and post-activity) was calculated for each learner who answered all questions within a given activity. A Cohen’s d was determined for each activity based on overall per cent correct score for these matched participants. Pooled effect sizes were determined by averaging Cohen’s d across activities. Benchmarks for educational effectiveness were determined from the top and bottom 25% of Cohen’s d across activities.

All data used in this analysis were devoid of personal identifiers and analysed in MS Excel.

Results

Overall, matched pre- and post-activity test score data were available for 5980 learners (average of 150 per activity), which represented 8.9% of overall unique participants accessing the activities. Eighty-two per cent of these matched participants were clinicians. The pooled Cohen’s d across all 40 activities was 0.63, with 50% of all Cohen’s d measures falling between 0.48 and 0.75.

There was an average of 154 matched participants per activity across 31 digital publications and 134 per activity across 9 webcasts. The pooled Cohen’s d was 0.67 for digital publications and 0.50 for webcasts. Pooled effect size was not substantively different for activities targeting primary care, specialists, or mixed primary care-specialist learners.

Pooled effect size across the 15 activities addressing infectious disease was 0.73; the pooled effect size for oncology-based activities (n = 5) was 0.40. There was an insufficient number of activities in other content areas to warrant pooling effect size.

The number of total and matched participants, pre- and post-activity test scores, standard deviations, and Cohen’s d for each of the 40 activities included in this study are detailed in the following Table 1.

Discussion

The pooled Cohen’s d across all 40 internet-based enduring materials was 0.63, with 50% of all Cohen’s d measures falling between 0.48 and 0.75. As benchmarks, Cohen [5] recommended the following: 0.2 (small effect), 0.5 (medium effect), and 0.8 (large effect). Accordingly, the pooled magnitude of effect for these 40 activities could be interpreted as “medium”. However, the question remains: Is medium good? In response, Cohen cautioned against using these benchmarks as anything other than preliminary. The expectation was for each industry to establish its own benchmarks based on repeated measurement. For CME providers new to effect size measurement, the first place to look for such benchmarks would be peer-reviewed literature.

A 2008 meta-analysis by Cook et al. [2] pooled 126 internet-based educational interventions designed to
Table 1. Standardised Mean Difference for 40 Internet-based Enduring Materials: 2016–2018.

| Content focus          | Format       | Target audience | Total unique participants | Total matched (pre/post) participants | Number of assessment questions | Pre-activity score (SD) | Post-activity score (SD) | Cohen's $d$ |
|------------------------|--------------|-----------------|---------------------------|---------------------------------------|-----------------------------|-----------------------|-------------------------|-------------|
| Oncology               | Digital publication | Specialist     | 609                       | 76                                    | 4                           | 66.5%                 | 85.9%                   | 0.7         |
| Immunology             | Digital publication | Specialist     | 655                       | 83                                    | 5                           | 50.0%                 | 64.0%                   | 0.52        |
| Genetics               | Digital publication | Specialist     | 721                       | 63                                    | 4                           | 39.3%                 | 64.3%                   | 0.84        |
| Musculoskeletal disorders | Webcast     | Specialist     | 729                       | 46                                    | 4                           | 35.9%                 | 51.1%                   | 0.66        |
| Infectious disease Oncology | Webcast   | Specialist     | 733                       | 49                                    | 4                           | 49.5%                 | 62.2%                   | 0.36        |
| Neurology              | Digital publication | Specialist     | 848                       | 80                                    | 4                           | 51.3%                 | 69.1%                   | 0.75        |
| Pulmonology            | Digital publication | Primary care   | 855                       | 104                                   | 4                           | 69.7%                 | 84.9%                   | 0.6         |
| Genetics               | Digital publication | Specialist     | 911                       | 51                                    | 5                           | 40.0%                 | 52.9%                   | 0.58        |
| Oncology               | Digital publication | Specialist     | 1035                      | 100                                   | 4                           | 36.0%                 | 47.0%                   | 0.38        |
| Cardiology             | Digital publication | Primary care   | 1037                      | 218                                   | 5                           | 29.9%                 | 43.4%                   | 0.56        |
| Infectious disease Immunology | Digital publication | Specialist     | 1040                      | 152                                   | 4                           | 53.8%                 | 78.0%                   | 0.88        |
| Infectious disease Oncology | Digital publication | Specialist     | 1077                      | 85                                    | 6                           | 61.6%                 | 74.9%                   | 0.64        |
| Infectious disease Infectious disease | Digital publication | Specialist     | 1096                      | 165                                   | 6                           | 45.7%                 | 68.0%                   | 0.98        |
| Oncology               | Digital publication | Specialist     | 1104                      | 148                                   | 6                           | 37.8%                 | 58.8%                   | 0.93        |
| Infectious disease Mixed primary care/ specialist | Digital publication | Mixed primary care/ specialist | 1172                      | 182                                   | 4                           | 52.9%                 | 65.5%                   | 0.55        |
| Gastroenterology       | Webcast      | Mixed primary care/ specialist | 1258                      | 201                                   | 4                           | 47.4%                 | 71.4%                   | 0.98        |
| Infectious disease Addiction medicine Dermatology | Digital publication | Primary care   | 1287                      | 104                                   | 6                           | 55.8%                 | 62.8%                   | 0.38        |
| Cardiology             | Digital publication | Specialist     | 1301                      | 281                                   | 6                           | 37.2%                 | 52.1%                   | 0.55        |
| Infectious disease Cardiology | Digital publication | Specialist     | 1373                      | 127                                   | 4                           | 56.4%                 | 73.8%                   | 0.72        |
| Infectious disease Oncology | Digital publication | Specialist     | 1459                      | 145                                   | 4                           | 31.0%                 | 42.6%                   | 0.55        |
| Infectious disease Mixed primary care/ specialist | Digital publication | Mixed primary care/ specialist | 1538                      | 155                                   | 4                           | 51.0%                 | 79.3%                   | 1.18        |
| Oncology               | Digital publication | Specialist     | 1589                      | 73                                    | 6                           | 54.1%                 | 71.0%                   | 0.63        |
| Pulmonology            | Digital publication | Specialist     | 1683                      | 33                                    | 6                           | 36.4%                 | 45.5%                   | 0.39        |
| Gastroenterology       | Digital publication | Specialist     | 1790                      | 93                                    | 6                           | 53.8%                 | 76.3%                   | 0.94        |
| Infectious disease Dermatology | Digital publication | Primary care   | 1975                      | 154                                   | 5                           | 50.0%                 | 61.0%                   | 0.46        |

(Continued)
Table 1. (Continued).

| Content focus | Format            | Target audience | Total unique participants | Total matched (pre/post) participants | Number of assessment questions | Pre-activity score (SD) | Post-activity score (SD) | Cohen’s $d$ |
|---------------|-------------------|-----------------|---------------------------|---------------------------------------|-------------------------------|-------------------------|--------------------------|-------------|
| Infectious disease | Webcast             | Specialist        | 2224                      | 142                                   | 5                             | 51.1%                  | 66.2%                     | 0.51        |
| Endocrinology | Digital publication | Primary care      | 2289                      | 201                                   | 4                             | 48.4%                  | 58.3%                     | 0.38        |
| Infectious disease | Digital publication | Primary care      | 2354                      | 98                                    | 6                             | 40.7%                  | 56.0%                     | 0.60        |
| Psychiatry    | Digital publication | Mixed primary care/specialist | 2698               | 346                                   | 4                             | 32.3%                  | 44.3%                     | 0.48        |
| Endocrinology | Digital publication | Primary care      | 3074                      | 630                                   | 5                             | 71.0%                  | 86.0%                     | 0.65        |
| Cardiology    | Webcast             | Specialist        | 3233                      | 74                                    | 6                             | 48.0%                  | 59.0%                     | 0.48        |
| Infectious disease | Webcast             | Primary care      | 3446                      | 298                                   | 6                             | 24.6%                  | 47.4%                     | 1.00        |
| Endocrinology | Webcast             | Mixed primary care/specialist | 3725               | 223                                   | 4                             | 61.1%                  | 70.5%                     | 0.39        |
| Infectious disease | Digital publication | Primary care      | 3885                      | 219                                   | 4                             | 35.2%                  | 48.6%                     | 0.50        |
| Infectious disease | Digital publication | Primary care      | 4589                      | 352                                   | 5                             | 23.1%                  | 39.2%                     | 0.76        |

Affect knowledge outcomes among clinicians. The pooled standardised mean difference (Hedges $g$) for these interventions was 1.0. As with the 40 internet-based enduring materials reported in this study, this pooled standardised mean difference was reported in comparison with no alternate intervention. Although Cook et al. used Hedges $g$, the result is directly comparable to Cohen’s $d$ with the exception of small sample sizes, which is not applicable here. Accordingly, the pooled effect size from our 40 studies (0.63) appears substantively lower than that reported by Cook et al. In that Cohen’s $d$ is a proportional metric, this comparison suggests that the activities included in the Cook et al. meta-analysis were nearly 60% more effective overall than the 40 activities described in this research. However, before such a CME provider should begin lamenting past efforts, the comparability of this new benchmark must be assessed.

In order to be considered directly comparable, effect size measures should be derived from similar interventions using similar techniques. In regard to the 40 activities described in this research, effect size was derived from two types of internet-based enduring materials (digital publication and webcast) targeting primary care and specialty clinicians (i.e. MD, DO, NP, PA) using a one-group (matched), pre- versus post-activity design. Effect sizes in the Cook et al. analysis were derived from three different designs: 1) one-group, pre- versus post-activity; 2) two-group (i.e. unmatched), pre- versus post-activity; and 3) two-group (i.e. participants and non-participant control group), post-test only. Interestingly, post-test only studies ($n = 25$) reported the lowest pooled effect (0.66) versus 0.88 for two-group, pre- versus post-activity ($n = 25$) and 1.18 for one-group, pre- versus post-activity ($n = 76$ studies). Although 1.18 may appear to be a more comparable benchmark for the 40 studies in this research given assessment design, the interventions and target learners from which it was derived likely present limitations. For example, the target learners were more diversified in the Cook et al. meta-analysis, with only 21% of participants including physicians in practice (the remainder was composed of medical students, physicians in post-graduate training, and nurses and pharmacists in practice and in school). Moreover, the interventions included in the meta-analysis were more varied in design (e.g. synchronous and asynchronous delivery, containing various levels of interactivity, and inclusion of online discussions and tutorials).

In a subsequent study of 114 internet-based CME activities targeting only physician participants, Casebeer et al. [3] reported a pooled standardised mean difference of 0.82. Effect sizes (Cohen’s $d$) in this report were derived from the two-group, post-test only design. Notably, this design was associated with a dampening of effect in Cook et al. [2]. Activities included in this analysis were described in three different formats: 1) interactive text-based (e.g. conference coverage, special reports, and basic clinical updates), 2)
interactive case-based (e.g. case-based polling questions with feedback), and 3) multimedia (e.g. live or roundtable presentations with video lectures). The greatest pooled effect was reported for multimedia activity (1.26) versus 1.08 for case-based and 0.58 for text-based. As with the 40 activities reported in this research, the activities reported in Casebeer et al. [3] addressed multiple content areas, most commonly psychiatry (21%), cardiovascular (16%), and infectious disease (11%).

In addition to studies by Cook et al. and Casebeer et al., there has been one additional meta-analysis reporting effect size for online learning; however, this report did not include knowledge-level outcomes[8].

**Conclusion**

This study describes the pooled effect size of 40 certified, internet-based enduring materials. Although the findings are not directly comparable with published literature, they are more likely to resemble the practice of CME providers responsible for producing nearly 50,000 such activities in 2018 alone. Based on this report, a benchmark Cohen’s d between 0.45 and 0.78 could be considered within the expected range for effectiveness. Activities with effect sizes below this range should be reconsidered for future programming, whereas activities with effect sizes above this range should be evaluated to clarify elements associated with success. There are still concerns, however, with this benchmark beyond how the activities were structured, the target audience, and the method of calculating effect size (e.g. one-group, pre- vs post-activity). For example, the median number of pre- and post-assessments included in each study was 4, but how do we know this number is sufficient? Moreover, guidelines from the National Board of Medical Examiners were followed in the development of assessments, but without any formal psychometric testing, how do we know these questions measured their intention with any reliability? Cook et al. reported that study quality was inversely proportional to effect, with more poorly designed studies reporting a higher effect size[2]. How do we know the degree to which such is in play with the 40 activities analysed in this research?

Overall, effect size measures can serve as useful benchmarks for CME providers; however, the current state of available data in this industry is likely insufficient. The data and cautions included within this research will ideally serve to help CME providers refine their future measurements.

**Disclosure Statement**

Both authors are employed by Med-IQ, LLC. There is no conflict of interest.

**References**

[1] Accreditation Council for Continuing Medical Education. ACCME date report: growth and advancement in continuing medical education—2018. [cited 2020 Jul 13]. Available from: http://www.accme.org/sites/default/files/2019-07/818_20190716_2018_Data_Report.pdf

[2] Cook DA, Levinson AJ, Garside S, et al. Internet-based learning in the health professions: a meta-analysis. JAMA. 2008;300(10):1181–1196.

[3] Casebeer L, Bown J, Roepke N, et al. Evidence-based choices of physicians: a comparative analysis of physicians participating in Internet CME and non-participants. BMC Med Educ. 2010 Jun;10(10):42.

[4] Olivieri J, Harris D, Dietze D, et al. A call for calculating effect size in CME outcome measurement. CE Meas. 2012;6:22–25.

[5] Cohen J. Statistical power analysis for behavioral sciences. 2nd ed. Hillsdale, N.J: Lawrence Earlbaum Associates; 1988.

[6] Accreditation Council for Continuing Medical Education. How is an internet enduring material activity defined? [cited 2020 Jul 13]. Available from: http://www.accme.org/faq/how-internet-enduring-material-activity-defined

[7] Case SM, Swanson DB. Constructing Written Test Questions for the Basic and Clinical Sciences. National Board of Medical Examiners, Philadelphia, PA; 1998.

[8] Mansour M, Lockyer J. A meta-analysis of continuing medical education effectiveness. J Contin Educ Health Prof. 2007;27(1):6–15.