ONLINE ASSESSMENT DESIGNS IN MOOCS:  
A META-ANALYSIS

Malissa Maria Mahmud  
Centre for English Language Studies, Sunway University  
Petaling Jaya, Selangor, Malaysia

Susila Madasamy  
Centre for English Language Studies, Sunway University  
Petaling Jaya, Selangor, Malaysia

Yazilmiwati Yaacob  
General Studies Department, Sunway University  
Selangor, Malaysia

Othman Ismail  
Academy of Language Studies, University of Technology MARA  
Selangor, Malaysia

Chandra Reka Ramachandiran  
School of Computing, Asia Pacific University  
chandra.reka@apu.edu.my

ABSTRACT

The impetus to advocate alternative modes of education through online platforms has begotten the media’s attention and to date, major research universities have instituted online content for worldwide consumption. Originally conceptualized as a means for escalating the reach and accessibility of learning materials, Massive Open Online Courses (MOOCs) have been seen as the forefront of the 21st century education. However, the development of MOOCs has stagnant over the recent years. This is especially the case for variability of assessments styles in MOOCs, vital in regulating the quality of education. Thus, this paper examines the natures of emerging assessment styles in MOOCs and their respective outcomes. Eight samples were derived based on their relevance and established priori to probe on the assessments’ designs in which the magnitude of effect size (ES) for the selected samples were tabulated according to the Cohen’s $d$ formula and benchmark (1988; 1992). The outcomes were organized according to four categories of variables: Performance, Attitude, Interaction and Satisfaction. Additional qualitative data was also recorded. Findings juxtapose that the majority of outcomes were within the Performance variable, measured with significant number of small effect sizes. Medium and large effect sizes were also yielded from multiple outcomes; however, the generated negative effect sizes were also noticeably present. Overall, the effect sizes suggest an all-encompassing positive repercussions of quality assessment employed within the MOOCs context. Nonetheless, this paper acknowledges the limitations and possible confounds within the current landscape of MOOCs assessment styles; thus, further researches are indispensable to grasp the complete functionality.

KEYWORDS

MOOCs, Assessment, Cohen’s $d$, Effect Size, Educational Technology.

1. INTRODUCTION

The contention of technological integration in the context of teaching and learning has indisputably gained immense and legitimate considerations in the past decade. The advances in technological interventions and Web 2.0 have impelled HEIs in Malaysia to experiment with novel pedagogical
models to support the online portion of instructions. With these developments, innovative trajectories are adopted by institutions to align and position themselves as being more competitive and relevant to the millennial learners. More significantly, the diffusion of technology in this circumstance – the blend of both conventional approach and technology – has rendered the “chalk and board” method rather irrelevant to the current teaching and learning context. It is interesting to observe that within the occurrence of blended learning, more fascinating, specific approaches and interventions have been introduced to facilitate teaching and learning; namely, collaborative learning, problem-based learning, student-centred or active learning, flipped classrooms and the latest trend – MOOCs. The academia has seen an advent of Massive Open Online Courses (MOOCs), claiming to be one of the game changers in enhancing and augmenting existing instructional strategies in the landscape of higher education. MOOCs began to receive huge attention in 2011 when prominent American universities instituted MOOCs through various avenues such as Coursera and Udacity (Milligan, Margaryan, & Littlejohn, 2013). The term was initially used to illustrate an online open course ‘Connectivism and Connective Knowledge (CCK08)’, which was established at the University of Manitoba by George Siemens and Stephen Downes. With 2200 participants enrolled worldwide, MOOCs are asserted to offer a myriad selection in regards to various disciplines, catering for an immense number of partakers from all over the globe who attend the online classes free of charge (Liyanagunawardena, Adams, & Williams, 2013). Moreover, MOOCs promotes lifelong learning in a way that is in line with many HEIs.

Nevertheless, the status quo of its implementation and feasibility are still at the experimental stage due to its perceived limitations. First and foremost, there are some who are convinced that the concept of reaching and catering to anyone and anywhere at any time is implausible (Yousef, Chatti, Schroeder, Wosnitza, & Jakobs, 2014). It is also argued that MOOCs defeat the purpose of student-centeredness approaches. Yousef, Chatti, Schroeder and Wosnitza (2015) stated that the employment of MOOCs is very much inclined towards the conventional approach wherein teachers are the central point and in control of the whole teaching and learning process. One of the major constraints in the implementation of MOOCs is the involvement of assessment and feedback essential to gauge students’ progress (Staubitz, Petrick, Bauer, Renz, & Meinel, 2016). The absence of communication and contact between learners due to the uploaded video may have contributed to the significantly-low retention rate. Yousef, Chatti, Schroeder and Wosnitza (2015) reported that the average dropout rate is as high as 95%. Deliberations on the factors contributing to such high rates revolve around the varying culture and motives among the enrolled learners (Yousef, Chatti, Schroeder, & Wosnitza, 2015).

Stakeholders in the education industry deem MOOCs as possessing a potential to be part of the higher education topography, a substitute method of teaching and learning. In view of the existing constraints and issues, further discourse to overcome these limitations has prompted a transferal of design archetypes and models to address the prevalent issues concerning MOOCs. With its growing popularity and endorsement from HEIs, the US quality benchmarking and assurance program stated that MOOCs have the prospect of offering quality education and being integrated into higher education programmes as they are designed for archetypal students (Legon, 2013). Nevertheless, concerns have been raised since this innovative approach lacks emphasis on high quality content, and mechanisms to gauge quality are still non-existent (Conole, 2013). As numerous universities worldwide are investing resources to establish MOOCs, it becomes a ubiquitous issue to solidify the quality-assurance mechanism before accounting it as a key principal part of its implementation.

Ubiquity of MOOCs has commanded the reconceptualization of traditional assessments. Scarce circumstantial evidence is inferred within the western context, with minimum grasp of how it would be realistically delivered in developing countries (Hyman, 2012; Pappano, 2012). Predominantly, MOOCs use peer-assessment to gauge the deployed intervention (Meek, Blakemore, & Marks, 2017). Many have argued that unvetted and unstructured peer assessment results deployed within the MOOCs environment lack credibility, and are overly reliant on multiple-choice quizzes (Suen, 2014; Xiong, Goins, Suen, Pun, & Zang, 2014). In addition, due to
the high enrolments in MOOCs, manual grading and feedback providing are not feasible. Albeit deemed unviable to an extent, there are existing assessment tools which have been leveraged among the practitioners to measure learning outcomes, specifically designed for both formative and summative assessments. Similar to any teaching and learning context, the assessment design in MOOC must promote online engagement via autonomous pedagogical means. Nonetheless, the enormous data generated by learners in a MOOC can be used for developing and refining automated assessment techniques. Perhaps most broadly considered, the rhetoric of how MOOCs can “overcome inequality” in terms of access and quality of higher education needs to be reconsidered in a number of ways. Considering the incongruity present, this paper aims to examine the natures of emerging assessment styles in MOOCs and their respective outcomes by employing meta-analysis.

2. RESEARCH METHOD

Meta-analysis or statistical means of coalescing samples from numerous studies was utilized for this study. The method was first developed by Smith and Glass in 1977, in which the valuations of magnitude or effect size are yielded, to compound statistical power, resolving ambiguous contexts when individual results seem to conflict. Examining novelties of MOOCs, specifically the assessments’ trends, can concede best practices.

For this study, a priori; inclusion and exclusion criteria were established before the screening process commenced. Published studies from Google Scholar, reference lists and various electronic databases spanning from the year 2015 to 2018 were perused. The databases - Wiley Online Library, Taylor & Francis Online, Springer, ERIC, Elsevier, ScienceDirect, ResearchGate, and Questia, which are prevalent with digital and technology-based publications, were shortlisted. Eligible studies in this research adhered to the established inclusion criteria consisting of a controlled (experimental or quasi-experimental) design, probe into MOOCs as the means of instructions and assessments and report on the student’s learning outcomes, which succinctly defined the study design, independent variables and dependent variables. At this juncture, 280 studies were excluded as they had failed to meet the stipulated criteria. In addition, samples must encompass keywords such as “MOOC”, “control”, “treatment” and “dependent measures”, “online”, “technology”, “computer” adapted from relevant meta-analysis studies (Sitzmann et al., 2006; Cook et al., 2008; Bernard et al., 2009; Means et al., 2013; Tamim et al., 2011).

As a result of the rigorous screening process, only eight samples were successfully included. The samples were subsequently analyzed to measure the effect sizes, where Cohen’s $d$ (1988; 1992) formula was applied. For samples that quantified the $t$ value, $F$ value, $p$ level and frequency, Formula 1 and Formula 2 (see below) were applied. Hall and Rosenthal (2018) stated that the ES is calculated where the mean variance between the experimental and control groups, while the denominator is the pooled standard deviation (PSD). A pooled value from both groups was used if that value was not specified. The benchmark by Cohen (1988) postulated the basis for interpreting the power of the computed ES, which are delineated as (a) "small, $d = .2$", (b) "medium, $d = .5$", and (c) "large, $d = .8$".

Formula 1: The ES is the standardized mean difference between two groups

$$d = \frac{\text{Mean}(\text{experimental}) - \text{Mean}(\text{control})}{\text{pooled Std. dev.}}$$

Formula 2: Cohen’s $d$ in relation to $t$-test is used as formula

$$d = \frac{t}{\sqrt{df}}$$
### 2.1 Findings and Discussion

Table 1. The selected 8 multi-outcome samples, interventions, outcomes, p-value and mean ES and feedback.

<table>
<thead>
<tr>
<th>No</th>
<th>Article/Sample</th>
<th>Assessment (Intervention/Tool)</th>
<th>Sample Size</th>
<th>Outcome</th>
<th>Variable</th>
<th>P-Value</th>
<th>Effect Size</th>
<th>Strength</th>
<th>Feedback (survey/focus group/interview)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ashton &amp; Davies, 2015</td>
<td>Scaffold Rubrics/ (xMOOCs)</td>
<td>5690</td>
<td>Student rating (means for each rubric item by writing sample): Overall</td>
<td>Performance</td>
<td>.295</td>
<td>.1055</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Experimental Group (E) Rubric-plus intervention - focusing raters on the pertinent aspects of the rubric tends to improve rating validity and reliability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Control Group (C) Rubric-only condition - did not include any construct explanation or rubric guidance to raters.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Student rating (means for each rubric item by writing sample): Sentence Variety</td>
<td>Performance</td>
<td>.163</td>
<td>.1286</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Student rating (means for each rubric item by writing sample): Hook</td>
<td>Performance</td>
<td>.002</td>
<td>.2672</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Student rating (means for each rubric item by writing sample): Blocking</td>
<td>Performance</td>
<td>.043</td>
<td>.1822</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Student rating (means for each rubric item by writing sample): Spelling</td>
<td>Performance</td>
<td>.432</td>
<td>.1172</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Chudziki, Pritchar d, &amp; Chen, 2015</td>
<td>Variation of practise problems (edX Learning Management System)</td>
<td>219</td>
<td>Correct rates on the problems they attempted: Between E1 and E2 Correct rates on the problems</td>
<td>Performance</td>
<td>.27</td>
<td>.1067</td>
<td>N/A</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>No</th>
<th>Article/ Sample</th>
<th>Assessment (Intervention/ Tool)</th>
<th>Sample Size</th>
<th>Outcome</th>
<th>Variable</th>
<th>P-Value</th>
<th>Effect Size</th>
<th>Strength</th>
<th>Feedback(survey/ focus group/ interview)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Gamage et al., 2017</td>
<td>Peer Identification and Aligned Incentives (Learning Management Systems - LMS)</td>
<td>85</td>
<td>Academic Performance: Between E1 and E2</td>
<td>Performance</td>
<td>.000</td>
<td>.8339</td>
<td>Large</td>
<td>Students’ motivation E1 - 53 E2 – 49 C – 29 E1 and E2 showed interest in further communicatio n in comparison to C Survey - 73% respondents said that being identifiable increases their interest in reviewing 69% respondents found the feedback useful 77% respondents are satisfied with the reviews</td>
</tr>
<tr>
<td>No</td>
<td>Article/Sample</td>
<td>Assessment (Intervention/Tool)</td>
<td>Sample Size</td>
<td>Outcome</td>
<td>Variable</td>
<td>P-Value</td>
<td>Effect Size</td>
<td>Strength</td>
<td>Feedback (survey/ focus group/ interview)</td>
</tr>
<tr>
<td>----</td>
<td>----------------</td>
<td>-----------------------------</td>
<td>-------------</td>
<td>---------</td>
<td>----------</td>
<td>---------</td>
<td>-------------</td>
<td>----------</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td>4</td>
<td>Huisman, Admiraal, Pilli, van de Ven, &amp; Saab, 2018</td>
<td>Peer Review on MOOCs (Coursera Platform)</td>
<td>26889</td>
<td></td>
<td>Correlation between 2 essay assignments</td>
<td>Performance</td>
<td>&lt; .001</td>
<td>.2765</td>
<td>Small</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Experimental Group 1 (E1)</td>
<td>565 participants grouping (DV only)</td>
<td></td>
<td>Correlation between 1st essay and quizzes</td>
<td>Performance</td>
<td>&lt; .001</td>
<td>.4889</td>
<td>Small</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Experimental Group 2 (E2)</td>
<td>565 participants grouping (DV only)</td>
<td></td>
<td>Peer reviewers’ ability and authors’ essay performance</td>
<td>Performance</td>
<td>.044</td>
<td>.013 (R^2)</td>
<td>Small</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Ability of intermediate authors and their essay performance</td>
<td>Performance</td>
<td>&lt; .001</td>
<td>.046 (R^2)</td>
<td>Small</td>
</tr>
<tr>
<td>5</td>
<td>Kelly &amp; Heffernan, 2016</td>
<td>Changing threshold of consecutively correct response (CCR) to optimise practice (Random assignment feature in ASSISTments)</td>
<td>412</td>
<td>Academic Performance</td>
<td>.89</td>
<td>E2 &amp; E1 0.461</td>
<td>Small</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Experimental Group (E1)</td>
<td>2 CCR were required to complete assignment</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Experimental Group (E2)</td>
<td>3 CCR were required to</td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.356</td>
<td></td>
<td></td>
<td></td>
<td>0.999</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>E3 &amp; E2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.633</td>
<td></td>
<td></td>
<td></td>
<td>E3 &amp; E4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>Article/Sample</td>
<td>Assessment (Intervention/Tool)</td>
<td>Sample Size</td>
<td>Outcome</td>
<td>Variable</td>
<td>P-Value</td>
<td>Effect Size</td>
<td>Strength</td>
<td>Feedback (survey/ focus group/ interview)</td>
</tr>
<tr>
<td>----</td>
<td>----------------</td>
<td>--------------------------------</td>
<td>-------------</td>
<td>---------</td>
<td>----------</td>
<td>---------</td>
<td>------------</td>
<td>----------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>7</td>
<td>Reilly et al, 2016</td>
<td>Automated Essay Scoring (AES) assignment (edX Learning Management System)</td>
<td>303</td>
<td>Total Rubric Score: Between students with English as their first language (EFL) and students with English as their second language (ESL) in E</td>
<td>Performance</td>
<td>&lt; .01</td>
<td>.4156</td>
<td>Small</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Experimental Group (E)</td>
<td></td>
<td></td>
<td>Students whose essays were calibrated with AES essay according to rubric</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Control Group (C)</td>
<td></td>
<td></td>
<td>Students whose essays were calibrated by instructors according to rubric</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Rushkin et al., 2017</td>
<td>Adaptive Learning and Assessment (HarvardX)</td>
<td>518</td>
<td>Academic Achievement Score: Posttest of E and C</td>
<td>Performance</td>
<td>.21</td>
<td>.17</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Experimental Group (E)</td>
<td></td>
<td></td>
<td>Use of HarvardX in Super-Earths and Life course to adaptively interact with</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Control Group (C)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The results are organized into Table 1. It consists of the selected 8 multi-outcome samples, their respective interventions, the total of 40 outcomes, tabulated p-value and mean ES as well as any additional feedback from the participants within each sample. The interventions reviewed includes Adaptive Learning and Assessment, AES, CCR, Peer Review, Peer Identification and Aligned Incentives, Scaffold Rubrics, and variation of practice problems. The outcomes, in particular, are further organized into four categories of variables, namely, performance, attitudes, interaction and satisfaction. Most samples tested on the performance variable (n = 27), followed by the attitude (n = 11), interaction (n=3) and lastly satisfaction (n =1) variables. By noting the frequency of these categories, it illustrates that performance, is the focal point of investigated MOOCs in determining the value of their interventions.

The outcomes mean ES are organized according to their strength. Out of the 40 outcomes, 15 yielded small ES values, five yielded medium ES values and six yielded large ES values. The
remainder outcomes yielded 11 negative ES values and eight insignificant ES values. Having small ES values does not necessarily mean the interventions were ineffective. On the contrary, this large and widespread pool of small ES across samples, interventions and outcomes demonstrates the overall potential of the MOOC assessment. Cohen (1988, p. 25) cautioned on the application of the terms like “small”, “medium” and “large”, employed out of context. A small ES may not necessarily denote that the designed intervention is bad; it simply means the positive effects of intervention were marginally observable between the treatment and control groups. A negative ES however, does indicate the control group outperformed the treatment group, suggesting that the intervention backfired.

The efficacy of respective interventions was diminished on the outcomes by the presence of confounding variables, therefore leading to a smaller effect size. Moreover, the medium and large ES values demonstrate the applicability in these interventions. Interestingly, these values were found for performance, attitude and interaction variables. This suggests an all-encompassing positive effect of quality MOOCs assessments on the outcomes of interest. However, these ES values were found only within two of the eight samples collected. Next, the negative ES values are apparent within three samples only with very specific outcomes. A negative ES indicates that the control group outperformed the experimental group on respective outcomes, with some on a significant p-value. This might suggest that the presence of limitations to the MOOCs interventions investigated that could be further explored. Mahmud (2018) proposes that the applications of interventions with negative effect sizes should be used with caution. A study by Hone and El Said (2016) suggested that MOOC’s instructor can be a vital determinant for student’s retention rate. This could be supplemented by the qualitative data on student motivation and satisfaction in Gamage and colleagues (2017) study obtained from the sample. Lastly, the insignificant ES values, excluding those attributed to pretest conditions, show negligible differences between slight changes in the MOOCs assessment interventions. For instance, Chudzicki, Pritchard, & Chen (2015) had an insignificant ES on their outcome between two of their experimental groups that differ on the presentation of the problems, rather than any changes in the functionality of the assessment. It may be inferred that in the grand scheme of MOOC assessments, how the assessment is presented may not be as important as it appears.

3. CONCLUSION

Overall, the assessment strategies involved in MOOCs can enhance the online education system, as demonstrated by the medium and large ES values. There is potential for growth and understanding of the assessments in MOOCs, as exemplified in the small ES values. However, there is a considerable number of negative ES that cannot be overlooked. It may result from the incomplete understanding of the interactions with other confounding variables, and insignificant ES values, noting the importance of an overall working assessment system. Thus, more research is needed to further identify the exact nature of the interventions and their effects on outcomes as well as to explore multiple teaching methodologies and assessments to overcome confounding variables and limitations. It is also important to note on the learning domains when designing any assessments in which they are mapped with learning outcomes to support both teaching and learning (Mahmud, Yaacob, Ramachandiran, Ching, & Ismail, 2019).

To sum up, the various instructional strategies that have budded into existence because of the onset of blended learning are a fascinating phenomenon to observe. Without a doubt, these strategies possess innate potential and power in the context of teaching and learning yet to be uncovered. In other words, it is only a matter of time before sufficient researches are conducted on the subject matter, and the strategies aforementioned can truly alter the landscape of Massive Open Online Courses.
REFERENCES


References for the Eight Samples


