Motivation and Emotion
https://doi.org/10.1007/s11031-022-09929-2

ORIGINAL PAPER

Off the mark: Repetitive marking undermines essay evaluations due to boredom

Sinan Erturk[1] · **Wijnand A. P. van Tilburg**[2] · **Eric R. Igou**[3]

Accepted: 24 January 2022

Abstract
Essay-style assessment is widespread in education. Nonetheless, research shows that this tool can suffer from low reliability and validity. We attribute this problem partly to the boredom that marking multiple essays causes. Specifically, we propose that boredom in markers is associated with systematically lower marks on essays. To test this, we asked participants ($N = 100$) with an undergraduate degree to mark essays. The majority of these participants had at least some experience with marking. After marking each essay, participants indicated how bored they were. We found an increase in boredom over time and that higher boredom was associated with lower marks. Furthermore, offering a marking rubric did not prevent this problematic impact of boredom. These findings have implications for the validity of essays as an assessment tool and raise concerns about repetitive marking practices in general.

Keywords Boredom · Negativity bias · Assessments · Marking · Education · Emotion

Introduction

Essays are used for assessment throughout education. This is not surprising, as there are some clear benefits to using essays in assessment, such as providing an opportunity for students to show that they can analyse, synthesise and communicate knowledge of a topic (Williams et al., 1991). Essay questions, being open-ended, can be particularly relevant to real-world applications of knowledge, especially compared to selecting the correct response from a given set (Hift, 2014). Students who were asked to complete a survey on essay questions reported that they could display a greater degree of knowledge through them; students, therefore, believed that essays gave them a better opportunity to demonstrate academic ability (Bird et al., 2019). Despite the widespread use of essays to assess students, it is well-established that this type of assessment is prone to subjectivity and bias (Schaefer, 2008; Slomp, 2012). Therefore, in order to improve the use of essays as a form of assessment, it is crucial to identify and account for potential sources of bias, for example, through altering assessment policy and recommendations. Our research seeks to ascertain if and how boredom is a source of bias in evaluating essays.

Essay evaluations: benefits and challenges

The use of essay-style assessment is widespread throughout lower and higher education (Wyatt-Smith & Klenowski, 2013). Typically, essay assessment involves assigning students one or more broad questions that are answered in a few words (e.g., MEQs; Feletti, 1980) or more extensively (Pepple et al., 2010). The popularity of this assessment tool could in part stem from its perceived ability to facilitate content reflective of 'deep learning' in students, showing more analytical or creative thinking than other formats (Biggs, 1988; Parmenter, 2009). Assessments do not merely serve to measure performance and can directly affect students' motivation for learning (Cauley & McMillan, 2010), influencing learning processes (Tempelaar et al., 2013). Accordingly, the importance of thoroughly examining the utility of assessment methods lies at the centre of educational concerns (Kibble, 2017), and the reliability of the assessment

✉ Sinan Erturk
erturk.sinan@yahoo.co.uk

[1] Psychology Department, Insititute of Psychiatry, Psychology and Neuroscience, King's College London, 16 De Crespigny Park, London, UK

[2] Psychology Department, University of Essex, Colchester, UK

[3] Psychology Department, University of Limerick, Limerick, Republic of Ireland

Published online: 31 January 2022

is critical (Van der Vleuten, 2016; Van der Vleuten & Schuwirth, 2005).

Reliability for essay marking is usually evaluated in the educational literature using inter-rater or intra-rater indexes (Brown, 2009).[1] Prior research into assessment at universities, particularly in medical subjects, has suggested that marks given to essays often do not meet an acceptable level of reliability (Bell, 1980; Caryl, 1999; Newstead, 2002). This inconsistency in marking is pervasive in university settings, even for more experienced markers, with inter-rater reliability sometimes being as low as an r value of 0.41 (Newstead, 2002), and has been described as 'fragmented' within departments (Bloxham et al., 2016; Ecclestone, 2001). As a case in point, a study examining the reliability of essay-based psychology exams over five years found that overall inter-rater reliability for essay marks ranged from 0.54 to 0.61 for cognitive psychology, with worse scores for neuropsychology, which had reliability scores ranging from 0.25 to 0.50 (Caryl, 1999). One study examined how this inconsistency impacted real mark allocations and found that 24% of students at the start of secondary school could have been classified with an inappropriate mark (Meadows & Billington, 2005).

Given the popularity of essay-style assessment, it is obviously important to consider the quality of this tool. Some potential causes for the relatively low reliability of essay-style assessments have been identified already, allowing for improvements to be proposed. For example, amongst its challenges are the potential for subjectivity in markers (e.g., Pepple et al., 2010), confounding the assessed topic knowledge with language skills (Ackerman & Smith, 1988), and the potential loss of breadth of knowledge being tested as depth increases (Samuels & Coffinberger, 2004).

One improvement that is designed to reduce the subjectivity of markers is the use of a rubric. Rubrics are common tools that are used to increase consistency in marking. Some research suggests that they provide a common ground by which essays can be marked, leading to more reliable evaluations (Hack, 2013). Rubrics are guides for assessing the quality of a specific type of work. As studies have shown, this tends to increase reliability in marking. For example, one study demonstrated that when marking English essays without the aid of a rubric, only 3 out of 10 markers showed consistency within their marks for essays of similar quality (intra-rater reliability); this increased to 9 out of 10 when a rubric was used (Kayapinar, 2014). This, however, was not the case for consistency between markers, for which there was a consistent lack of reliability. A different study has more clearly shown that marking is more consistent when a rubric is used, finding an inter-rater reliability of 0.87 with full-time markers (Yang, 1987). However, some rubrics are more effective than others, and the way they are used is critical. For example, in one study, markers gave an average score for citations when that aspect was objectively poor, reflecting their overall view of the essay rather than a specific aspect as designated in the rubric (Reazai & Lovorn, 2010). This highlights the need to address factors that may affect the overall impression of the essay and account for them in our design.

Boredom and the marking of essays

One recurring and striking finding in the literature on essay marking is that as markers progressively mark more essays, they give lower marks (Bell, 1980; De Moira et al., 2002). Plausibly, marking multiple essays in a row is a monotonous task, requiring sustained attention. Such tasks tend to be associated with high levels of boredom (Ralph et al., 2017; Shackleton, 1981). This negative experience may contaminate markers' impression of essays, thus contributing to negative marking with an increasing number of essays that are marked. The reader might not be surprised to see boredom proposed as a potential source of bias in essay marking but may be surprised, however, that this has not been tested as of yet. To the authors' knowledge, the closest evidence of boredom's negative impact comes from the suggested detrimental impact of 'mental fatigue' in general (Klein, 2002; Klein & El, 2010). Indeed, it is well-established in medical education research that a prolonged cognitive load can lead to mental fatigue effects, resulting in worsening marks over time (McLaughlin et al., 2009). Specifically, this line of work on mental fatigue and cognitive load suggests that it becomes more difficult to engage in a cognitive task such as marking (Mizuno et al., 2011); this phenomenon would lead to less accurate marking. Interestingly, while mental fatigue is plausibly a correlate or perhaps characteristic of boredom, boredom tends to feature an unengaged, rather than cognitively loaded, mind (Eastwood et al., 2012). Our current research, in which we examined boredom in particular, is thus novel in its endeavour to examine how this state might negatively impact marks as it increases with prolonged marking.

What is boredom? Boredom is an unpleasant emotion (Smith & Ellsworth, 1987; Van Tilburg & Igou, 2017). It is characterised by lapses in attention and disengagement from tasks (Eastwood et al., 2012), feelings of being unchallenged, and the perception that the current activity (or lack thereof) is meaningless (Van Tilburg & Igou, 2012). It usually arouses a search for meaningful engagement (Van Tilburg & Igou, 2016), distraction (Moynihan et al., 2015), or impulsive action (Moynihan et al., 2017).

[1] Inter-rater reliability refers to the similarity in mark given by different individuals for the same essay. Intra-rater reliability refers to the similarity in marks by the same individual for similar quality essays.

Why and how might boredom affect marking? According to the 'affect-as-information' hypothesis, people's momentary affect at the time of making judgements can impact the judgement itself (Schwarz & Clore, 1983; Schwarz, 2012; Wyer et al., 1999; for an overview, see Martin & Clore, 2001). More generally, this ties into the 'mood congruency effect', which involves a negative affect tainting the evaluation of a target as congruent with the affective experience (Schwarz, 2012). For example, experiencing negative affect is associated with subsequently making more negative evaluations of people (Mano, 1992). As another example, Adaval (2003) found that the evaluation of brands became more negative when people had been induced with negative affect. Furthermore, the evaluation of employees in a work setting has also shown to be open to mood congruency effects, with a more negative mood resulting in more negative employee evaluation (Ding & Beaulieu, 2011). While general negative emotions such as anger, sadness, and fear are well documented in the literature, in terms of mood congruency effects on specific targets, boredom is an understudied emotion. Yet, within a marking context, the potentially negative impact of boredom, if it indeed exists, is important to identify.

To be clear, we suggest that aside from specific essays perhaps receiving lower marks due to them being boring, the marking process *itself* might cause boredom, which then reduces marks awarded. In other words, while any dull essay might (perhaps deservingly) receive a lower mark, we propose that the activity of repetitive marking of essays will, over time, culminate in such levels of boredom that evaluations, in general, will suffer, above and beyond how boring any individual essay might be. Note that we do not propose that boredom is the only factor potentially contributing to lower marks. Research suggests that other sources of bias exist, such as author gender, ethnicity, and physical attractiveness (Malouff & Thorsteinsson, 2016). Rather, we propose that boredom may well be another important, yet surprisingly overlooked, potential source of bias to examine.

As discussed earlier, a rubric is often employed to reduce inconsistencies between markers of essays. While some rubrics have shown to be effective in this endeavour, the results for the effectiveness of rubrics are somewhat mixed. Given the rather limited clear evidence for the effectiveness of rubrics in marking, we examined the presence of a rubric as an interesting but ultimately exploratory factor.

Research questions and hypotheses: an overview

This study puts forward boredom as a potential source of bias in the evaluation of written assessment. Given limited resources, markers must often mark large numbers of essays while meeting specific deadlines and fulfilling other aspects of their profession (Bloxham, 2009; Ecclestone, 2001; Smith & Coombe, 2006).

H1 Boredom increases with marking. Consecutive marking of multiple essays will increase boredom in the marker, reflected in a gradual increase in boredom as more essays are marked consecutively.

H2 Boredom lowers marks. Based on research on affect congruency effects in judgments, we propose an association between boredom and the marks assigned to essays. In particular, we hypothesise that as levels of boredom increase, essays will be evaluated more negatively.

H3 Increased marking leads to lower marks via increased boredom. It is reasonable to expect that essays that are themselves considered to be boring receive lower marks. However, we were primarily interested in the impact of boredom that resulted from the marking process rather than the boredom specific to any single piece of assessment. Specifically, we propose that boredom associated with the repetitive marking process itself is enough to compel markers into assigning lower marks. Therefore, we predicted that the more one has been marking, the more bored they are, which *in turn*, is associated with assigning lower marks. Put otherwise, the presumed positive association between boredom and time is responsible for lower marks, equivalent to a statistical indirect effect (Hayes, 2009) and requiring a partial association between boredom and assigned marks (after controlling for time spent marking) in addition to the total effect postulated under H2.[2] In addition to the above, we explored if a marking rubric might mitigate boredom's detrimental impact.

Methods

Participants and design

Participants were 102 people recruited from the crowdsourcing platform 'Prolific' (www.prolific.ac). We restricted our sample to people living in the United Kingdom who had achieved at least undergraduate level university education. This selection criterion ensured familiarity with marking procedures in this country and likely experience with academic essay writing. Two participants quit the study prematurely and were therefore excluded, resulting in a final sample of 100 participants ($M_{age} = 38.90$, $SD_{age} = 10.95$; 66 women, 34 men). All participants confirmed appropriate

[2] We also measured and tested how boredom and marks related to a lack of interest. The analyses and results are available on request.

English language ability, and 79 of them indicated that they had prior experience with essay marking.

Participants were randomly assigned to one of two conditions of a between-subject design: evaluating essays with or without the aid of a marking rubric. Each participant evaluated 10 essays in random order and of varying quality. The study required participants to finish within 1 h in total, but participants were informed at the start of the experiment that they should expect the study to take roughly 20 min. We gave participants £2 as an incentive for completing the study, and we received ethical approval from the Research Ethics Office at King's College London (MRS-18/19-8632).

Materials and procedure

Participants gave consent and then reported demographics, English language proficiency (1 = *poor*, 4 = *excellent*), and familiarity with essay marking (*yes* vs. *no*).[3] Next, we informed participants that they would be asked to evaluate 10 essays and to consider, in particular, "quality of writing, with a focus on the spelling, grammar and logical coherence of the piece." We instructed participants to give each essay a mark out of 100 (the percentage mark) and a letter grade ranging from A to F (the grade). For clarification, we outlined that essays given an 'A' should be "the best possible quality" and for 'F' essays of "unacceptable quality." To encourage participants to be accurate in their evaluations, we offered a £10 bonus on top of their regular participation payment for the three participants whose marks were closest to the average marks for their essays. We presented participants with examples of the different types of mistakes participants should look out for in essays (see Supplement). All participants also received a table that displayed the range of percentage values associated with each grade (e.g., 'A' corresponds to 85–100; see Supplement). Those in the rubric condition furthermore received descriptions and an approximate number of errors allowed for each grade to aid in their evaluation (see Supplement).

Participants then proceeded with the essay evaluations. Each essay was retrieved from the 'All-Essay' page on 'Blogspot' (http://all-essay.blogspot.com/). This educational website makes available short essays as learning tools for those studying English as a foreign language. We considered essays of 200 to 300 words in length, avoiding those that were highly culturally specific and ensuring that no single topic was covered more than once. From this set of essays, we randomly selected ten to feature in the study; selected essays covered topics such as the atmosphere, sharks, Christmas trees, and books. These essays were then carefully proofread, and obvious mistakes were corrected (e.g., spelling, grammar).

We created ten versions of each of these ten essays by systematically introducing zero through nine mistakes in each essay. These mistakes were each introduced in a randomly drawn sentence, with a maximum of one mistake per sentence. There were six types of mistakes introduced: incorrect tense verbs, sentences with some words jumbled in an incoherent order, made-up words, cases of incorrect punctuation, spelling errors, and words with the wrong meaning used. Artificially introducing these variations allowed us to (a) verify that participants indeed differentiated between comparatively high- and low-quality essays, and (b) ensured variation across the grading scale. While implementing a number of mistakes in essays does not reflect the variety of quality in real cases of marking, it is very useful in generating an objective measure of essay quality. This was selected as it avoided the issue with the subjective nature of quality that would be inherent to more realistic variations (see Supplement for examples).

Participants each evaluated the ten essays in a randomised order, and with a random number of mistakes. Participants were presented with the essay and a table. This table was the same one shown to them in the briefing. For those in the 'no rubric' condition, this was a table demonstrating the range of percentages associated with each grade. Those in the 'rubric' condition saw the table with full descriptions of what constituted an appropriate essay for each grade. Participants then entered the percentage mark they would give the essay in a provided box and indicated the grade on a slider that displayed a graphic of the letter they had selected. After they had confirmed their percentage and grade for that essay, they were asked "How bored do you feel at the moment?" and then they indicated how bored they were on a scale ranging from 1, "not bored at all", to 7, "extremely bored". They also indicated how interesting they had found the essay on another interval scale, ranging from 1, "very uninteresting", to 7, "very interesting". Participants proceeded to the next essay after this until they had evaluated all ten essays.

We then debriefed participants and thanked them for their time. It was made clear at this point that the study was examining the effect of boredom on essay evaluations and that they had been assigned to either one of two different conditions. They were told that they could direct any further enquiries towards the researcher.

Results

We analysed our data using SPSS statistics (IBM Corp, 2020) and *R* (R Core Team, 2017). First, we verified that the mistakes present in the essays indeed caused lower marks. In our study, each participant evaluated 10 essays. The marks

[3] Participants also completed a short big-5 personality measure and the short boredom proneness scale for exploratory purposes.

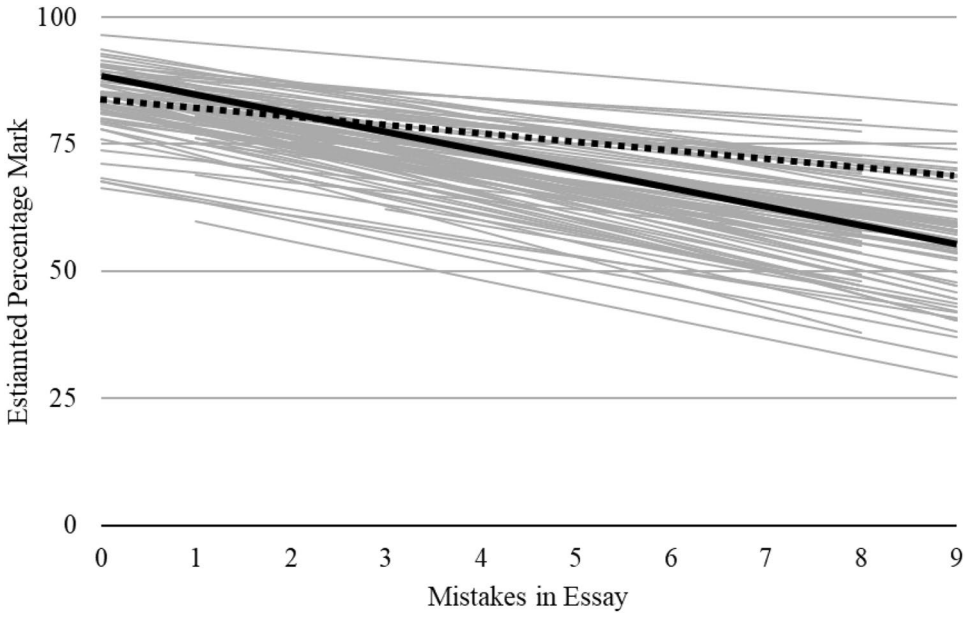

Fig. 1 Random slope analysis results for percentage mark by mistakes and rubric. Grey lines correspond to estimated marginal regressions for individual participants. Black lines represent estimated marginal means in the presence (solid) versus absence (dotted) of a marking rubric

thus represented both differences between participants (participants may differ overall in their average marks) and differences between individual essays marked by a participant (relative to a participant's average mark, an individual essay may score higher or lower). We analysed these nested data using a multilevel analysis to partition error terms separately for the 'higher' participant-level and 'lower' essay-level. We used maximum likelihood estimation here and in the other analyses unless stated otherwise. Percentage mark assigned to a specific essay by a specific participant was the dependent variable. The number of mistakes in the essay, the rubric condition, and their interaction were added as fixed-effect predictors (Fig. 1). We furthermore included a random-intercept to represent between-participants variance and a random-slope for the number of mistakes, given that the strength of the association between this variable and essay evaluations might vary between participants.[4] The current analysis, and other multilevel analyses, assumed an unstructured covariance matrix. The intraclass correlation for percentage marks, estimated with the empty multilevel model, was small but not negligible, $\rho = 0.204$.

This analysis revealed a (marginally) significant rubric × number of mistakes interaction, $F(1, 982) = 3.847$, $p = 0.050$, a significant main effect of number of mistakes, $F(1, 982) = 237.624$, $p < 0.001$, and no main effect of rubric, $F(1, 982) = 3.288$, $p = 0.070$ (Fig. 1). Participants who marked essays without the help of a rubric assigned approximately 2.7 percent points lower for each additional (artificial) mistake in an essay, $B = -2.718$, $SE = 0.279$, $95\%CI = [-3.266, -2.170]$, $t(982) = 9.736$, $p < 0.001$. In the presence of a rubric, this association was marginally more negative—an approximately 3.5 points lower percentage mark per mistake—$B = -3.510$, $SE = 0.292$, $95\%CI = [-4.084, -2.937]$, $t(982) = 12.018$, $p < 0.001$. Participants evaluated essays with mistakes more negatively, and those helped by a marking rubric perhaps did so using a slightly steeper marking curve.

Hypothesis 1: boredom increases with marking

The intraclass correlation for boredom, estimated with the empty multilevel model, was substantial, $\rho = 0.634$. Felt boredom was entered as the dependent variable in a random-slope multilevel analysis with time (equivalent to the position of essay in the sequence, i.e., essay 1 through 10), rubric, and the time × rubric interaction predictors of boredom. We added a random intercept for 'participant' and a random slope for 'time'. A significant main effect of time, $F(1, 991) = 228.800$, $p < 0.001$, revealed a significant positive association between this variable and boredom, indicating that boredom grew by approximately just under a tenth of a scale point for each additional essay marked, $B = 0.087$, $SE = 0.016$, $95\%CI = [0.055, 0.119]$, $t(991) = 5.367$, $p < 0.001$. Neither the main effect of rubric, $F(1, 991) = 1.384$, $p = 0.240$, nor the time × rubric interaction, $F(1, 991) = 1.649$, $p = 0.199$, was significant. As

[4] Results for the analyses of letter grades were similar to those of percentage grades. Details are available on request.

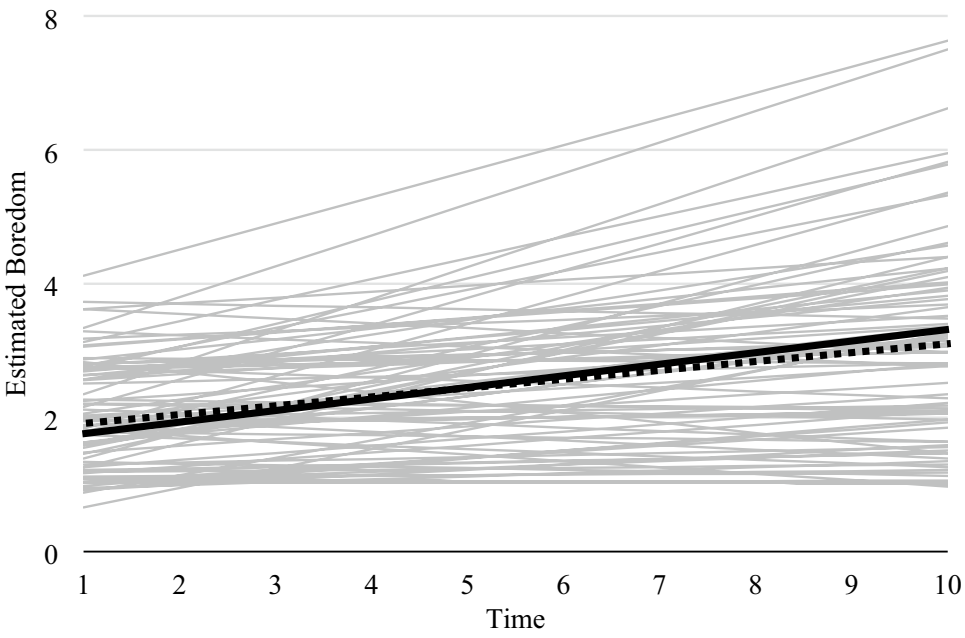

Fig. 2 Random slope analysis results for boredom over time by rubric. Grey lines correspond to estimated marginal regressions for individual participants. Black lines represent estimated marginal means in the presence (solid) versus absence (dotted) of a marking rubric

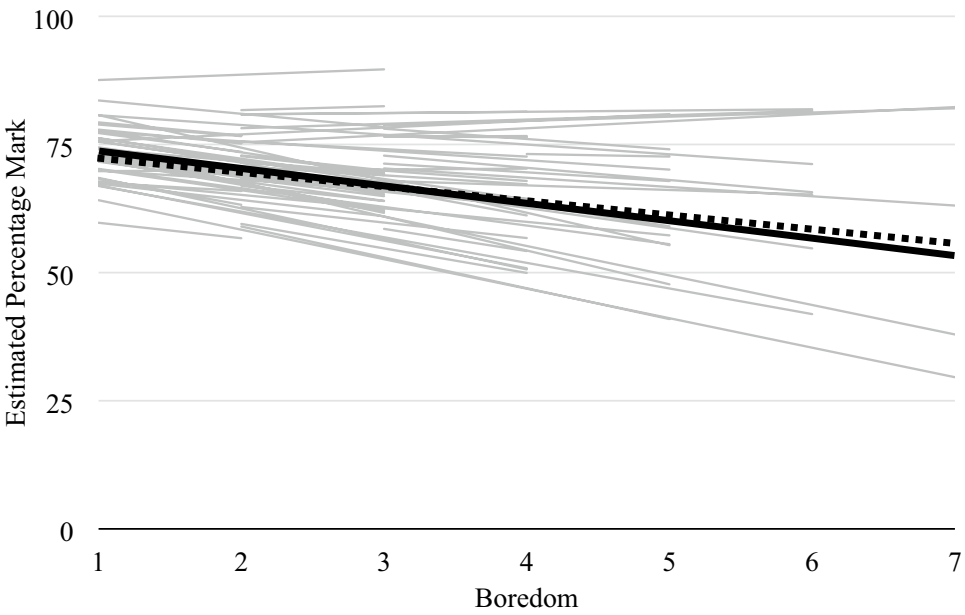

Fig. 3 Random slope analysis results for percentage mark by boredom and rubric. Grey lines correspond to estimated marginal regressions for individual participants. Black lines represent estimated marginal means in the presence (solid) versus absence (dotted) of a marking rubric. Individual participants' regression lines are not extrapolated beyond the range of their actual levels of reported boredom

participants marked more essays, their boredom gradually intensified (Fig. 2).[5]

Hypothesis 2: boredom lowers marks

To test if participants assigned lower marks as they became more bored, we entered the percentage mark as the dependent variable in a multilevel analysis with boredom as the predictor as well as the rubric and their interaction. We included a random intercept for participants, as before, and we also assigned a random slope to boredom. We found a significant main effect of boredom, $F(1, 977) = 17.009$, $p < 0.001$, indicating that, indeed, marks diminished as boredom mounted, $B = -2.845$, $SE = 0.690$, $95\%CI = [-4.199, -1.491]$, $t(977) = -4.124$, $p < 0.001$; each additional scale point of boredom came with a 2.8 point reduction in percentage mark assigned. Furthermore, neither the main effect of rubric, $F(1, 977) = 0.489$, $p = 0.484$, nor the boredom × rubric interaction, $F(1, 977) = 0.270$, $p = 0.604$,

[5] We report in the Supplement an exploratory analysis with also a quadratic term of time, which did not hold a significant association with boredom.

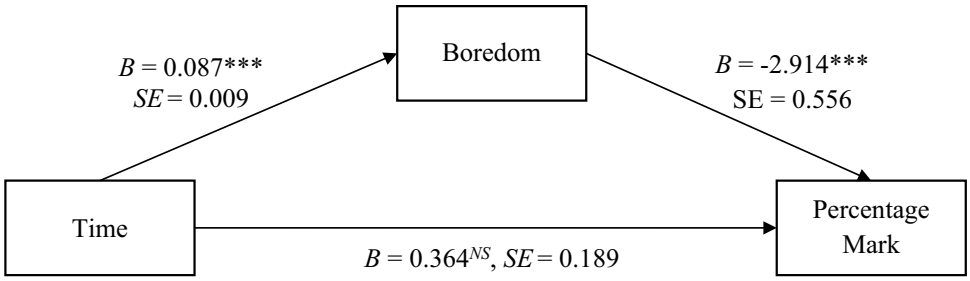

Fig. 4 Mediation sequence with time, boredom, and percentage mark. ***$p < 0.001$, $^{NS}p > 0.50$. Indirect association between time and percentage mark through boredom: $B = -0.254$, $95\%CI = [-0.367, -0.149]$

was significant (Fig. 3). These results suggest that, indeed, boredom diminishes marks given to essays, and this occurs regardless of the presence or absence of a marking rubric.[6]

Hypothesis 3: increased marking leads to lower marks via boredom

The previous analysis showed that marks tend to be lower when markers are bored. While we asked markers to report their overall level of boredom and not how boring any essay itself was, it is, of course, possible that the negative association between boredom and percentage marks simply shows that boring essays receive (and perhaps rightly so) lower marks. To test if the negative impact of boredom on marks is at least partly due to markers' gaining boredom over time, we examined if an indirect effect existed where time predicted *through boredom* lower marks (Fig. 4).

We estimated the first constituent direct association that makes up this hypothesized indirect association, the associating between time and boredom, using a random-intercept in which time served as a fixed predictor of boredom, with a random-intercept assigned to participants.[7] Again, boredom increased with time by slightly less than one-tenth a scale point for each additional marked essay, $B = 0.087$, $SE = 0.009$, $95\%CI = [0.070, 0.104]$, $t(993) = 10.021$, $p < 0.001$. Next, we tested the second constituent path: boredom's partial association with percentage marks controlling for time. We ran a random-intercept multilevel analysis with boredom and time as fixed predictors of percentage marks, alongside a random-intercept for participants. This analysis produced a significant negative partial association between boredom and the marks given to the essays, $B = -2.914$, $SE = 0.556$, $95\%CI = [-4.005, -1.824]$, $t(978) = 5.245$, $p < 0.001$, indicating that each unit of boredom came with an approximately 3 points lower percentage mark. The partial association between time and mark was not significant, $B = 0.364$, $SE = 0.189$, $95\%CI = [-0.006, 0.734]$, $t(978) = 1.929$, $p = 0.054$. We then used the Monte Carlo tool by Selig and Preacher (2008) to estimate the indirect effect, which was significantly negative, $B = -0.254$, $95\%CI = [-0.367, -0.149]$. These results suggest that the part of boredom that increases with time is at least partly responsible for a reduction in marks of approximately a quarter of a percentage mark. These results suggest that increasing boredom with marking over time appears to worsen marks awarded to essays, consistent with (H3).

Discussion

Overall, our findings suggest that as more essays are marked, boredom increases progressively in the marker. In line with this, high levels of boredom are associated with low marks given to the essay. More specifically, boredom increases as more essays have been marked, and the higher these levels of boredom, the lower the marks given to the essays.

Boredom increases with marking

Our results suggest that, as hypothesised, boredom increases as more essays are marked. This fits with previous research in the sense that feelings of boredom do tend to emerge from a monotonous task (Ralph et al., 2017). Given that markers typically report that the marking of essays is a boring task, this comes as no surprise (Schaefer, 2008). However, this is the first piece of research known to the authors that provides quantitative evidence for this. Notably, the essays used in this study were relatively short compared to usual marking environments, and it is probably not unusual that more than ten essays need to be marked by each marker. The finding that the increase in boredom over time was still found despite this study presenting participants with a shorter, less

[6] The supplement contains an additional exploratory analysis in which we test if the negative association between objective number of errors and marks was attenuated by boredom, which might suggest that boredom causes (also) *inaccuracy* in marking. The results did not reveal such an interaction.

[7] We excluded the random-slope terms for this mediation analysis to facilitate estimation of the indirect effect using the Monte Carlo tool by Selig and Preacher (2008). With inclusion of random slopes the direct effect are as follows: time to boredom, $B = 0.087$, $SE = 0.016$, $95\%CI = [0.055, 0.119]$, $t(993) = 5.322$, $p < 0.001$, boredom to percentage mark, (partial) $B = -3.432$, $SE = 0.650$, $95\%CI = [-4.709, -2.156]$, $t(978) = 5.277$, $p < 0.001$, and time to percentage mark, (partial) $B = 0.379$, $SE = 0.209$, $95\%CI = [-0.031, 0.789]$, $t(978) = 1.815$, $p = 0.070$.

demanding version of the task suggests that boredom may be a prominent issue in marking contexts. While this study was limited by time and resources, future research may examine the strength of this association with a more intensive task closer resembling real instances of essay marking.

Given that marking essays did increase boredom over time, we could then examine our other two hypotheses.

Boredom lowers marks

Our findings work in tandem with many examples of the mood congruency effect. Boredom's role as a negative emotion works with the idea that negative emotional states can result in negative evaluations of a specific target (Clore & Huntsinger, 2007). Effects such as these have been underreported in the literature on boredom. Furthermore, this study suggests that the observed lack of reliability in essay marking may be due to currently unaccounted boredom levels. This is key, as there are currently no measures to control the order in which essays are marked or how many essays are marked in one sitting. As a result, boredom may indeed explain the reduction in reliability and validity when essays are marked. Further research can examine whether targeting boredom in the marking of essays can increase the reliability and validity of the marking procedure.

The findings of this study more precisely suggest that boredom increases the more marking is done and that the greater the feeling of boredom, the lower the marks. This finding is consistent with previous research that shows that people give lower marks as they mark more essays (Bell, 1980; De Moira et al., 2002). As such, our study sheds light on boredom as a potential explanation for this phenomenon. Perhaps staggered marking or otherwise marking fewer essays in one sitting may help make evaluations more reliable.

We additionally found that the presence of a rubric did not prevent boredom from worsening evaluations. While we did not formulate specific hypotheses but rather treated the rubric as an exploratory variable, these findings add to the growing body of research on rubrics by suggesting that its impact may not be guaranteed or large. Of course, we did not examine how useful participants considered the rubric when evaluating essays or the degree to which they used them. Perhaps rubrics are used less when markers become more bored. As such, we can only conclude that rubrics were not sufficient to reduce the impact of boredom, not the reasons why this is the case.

Our research focused on the impact of general boredom on marking. Certainly, boredom is characterized by several other features. For example, boredom features mild negative affect, low or mixed arousal, thoughts about the purposelessness of a task, floundering attention to the task at hand, and a motivation to do something more challenging or meaningful (Danckert et al., 2018; Hunter & Eastwood, 2018). One of boredom's possible characteristics, or at least a close correlate, is mental fatigue (Gawron et al., 2001; Klein, 2002; Klein & El, 2010; Thompson et al., 2020). It is possible that mental fatigue, or some other boredom features, are more responsible for the impact that boredom has on marks then other features. The current research did not address boredom influence at the level of its specific features but investigating this further may offer insights into what drives boredom impact more specifically.

Limitations

One limitation of our study is the amount of variation in the essays. Firstly, they were on different topics, which is an unusual situation in essay marking; a marker would typically evaluate essays on a similar topic. Research suggests that if our stimuli were more similar, it would induce an even greater effect of boredom as it would be an even more monotonous task (Shackleton, 1981). Indeed, with a rapid drop in the novelty of the essays as more are marked, being more typical of essay marking, we would expect exaggerated effects with more realistically similar essay topics. We cannot, however, specifically conclude that this is the case. Therefore, it is important to ensure that we can replicate these findings with essays on a similar topic before applying them to essay marking in education. Similarly, future research should try to be more similar to real cases with longer essays. Once again, we would expect stronger associations in such a case. Also, these different topics could have resonated differently with certain individuals as well. For example, one essay discussed the atmosphere, which mentioned the damage to it caused by humans. To some, this is a topic they might be particularly passionate about. We attempted to control for this by having most essays be relatively neutral in terms of topic and mainly informative, but we cannot be sure that this was sufficient.

In terms of ecological validity, the mistakes we introduced cannot be said to match up exactly to the typical marking of essays. Usually, essays are evaluated on a more 'qualitative' sense of the content rather than a 'quantitative' number of mistakes. We introduced a quantitative measure of quality in order to tightly control for it, rather than equate it to reality. However, it may be the case that our measure of quality being quantitative made it inherently more boring to evaluate; indeed 'hunting for mistakes' seems less engaging than simply rating the essay on its content. Further research is therefore needed to see whether our observed effect persists with more ecologically valid assessment criteria.

Another limitation was the fact that our sample consisted of 'non-teachers'. While all participants had at least undergraduate degrees and thus an appreciation for the marking of essays, they are not the people who typically do this. Given

that this is not the case, our findings are not necessarily applicable to the education system with experienced markers. However, our intention was not to immediately expose a flaw with people who are marking professionally. Given that this is the first piece of research identifying worsening marks as a result of boredom, it was important to focus on the general phenomenon rather than its applicability in the real world. Furthermore, in our experiment, the markers had no way of knowing who was writing the essay and indeed no personal connection with the writer, this is not the case for teachers and students. Teachers may be able to identify writers from their essay and feel a sense of attachment knowing the essay is written by a student of theirs. This has implications for both boredom and other types of bias that our study does not go into. While there are issues when comparing our study to real life examples of essay evaluation, this does not undermine the key finding that this bias exists. It may be the case that in a more ecologically valid context, this bias is mitigated in some way, but regardless, this study suggests that this bias exists. This in fact opens up the idea of research with more experienced markers to ascertain whether they are still affected by this bias, and if not, how they achieve this. Findings from future research could help early-career markers address this bias before they gain sufficient experience to mitigate it.

As well as experience, it is worth considering that there are other moderators of this association that vary between individuals. Some people will be more cognitively engaged from marking than others, this would serve to mitigate the increase in boredom over time. Attention is another key aspect that can impact boredom, people with low attention control can become bored more quickly and vice versa (Westgate & Wilson, 2018). While our study did not measure such moderators, it would be valuable to examine whether and how these variables moderate the impact of marking on boredom.

Furthermore, it is important to bear in mind that our study addresses 'associations' between boredom and negative, or less accurate marks, not necessarily providing evidence that boredom is 'causing' these changes in marks. This is important when considering the congruency effect of boredom: that as a negative emotion, it negatively impacts evaluations made while experiencing it. Our study does not provide evidence of such a congruency effect, but future research may be able to address this by using boredom as an independent rather than dependent variable, comparing boredom to a control condition. It is challenging, however, to design a condition that does not induce boredom. It would be beneficial to perform a study where boredom is primed before they evaluate a study and perhaps compared to a different emotional prime instead of a non-boredom condition. However, the association found in this study does tell us that there is a relationship between increased boredom and both worse marks and less accurate marks. The study does not allow us to examine the mechanisms underlying this relationship. Also, our study does not allow us to investigate other potential factors that could influence essay marks other than boredom, such as the environment they are marking in, or how tired the marker is, or their general mood at the time of marking which could all influence their focus on the essay. These factors could all influence marks independent of boredom and would be worth considering in future research.

Given our finding that boredom from marking reduced interest in the essays, our study might point to a 'boredom congruency effect' in judgements. We can, of course, not draw that conclusion because our measure was whether participants found the target 'interesting' and not 'boring'. Even considering this limitation, our research might contribute to this very interesting, but understudied phenomenon. Further, we cannot clearly explain the negative congruency effects we observed with the common theories explaining such congruency effects. Perhaps, boredom served as a prime (e.g., Forgas, 1995), or it informed participants about the characteristics of the target (e.g., Schwarz & Clore, 1983). Future research needs to examine the specific processes underlying congruency effects based on boredom more closely.

Despite limitations in pinpointing the exact mechanism of the relationship between boredom and lower marks, this study highlights that the issue of marking a large number of essays infringes on the validity of later marks. As such, the requested length for essays should be carefully considered to ensure they are not unnecessarily long and thus straining the marker without good reason. Furthermore, this may suggest essays that are marked later could be worse off, so any constant order effects in terms of essay marking should be more tightly controlled. Simply being aware of this effect should allow measures to be implemented to prevent an impact. Such measures, such as taking regular breaks in between periods of marking, being given essays to mark in batches rather than all at once or dividing marking between more people should appeal to markers as well as mitigate the impact of boredom on their evaluations. Research suggests that boredom may be alleviated by finding meaning (Van Tilburg et al., 2013; Westgate & Wilson, 2018). To the extent that such a meaning search does not interfere with the task at hand, markers may attempt to remind themselves of the meaningful goals that essay marking ultimately serves (e.g., contributing to people's education). Furthermore, research shows that individuals turn to food in an attempt to alleviate boredom. While unhealthy snacking may be undesirable, 'exciting' healthy snacks such as miniature vegetables may provide an agreeable alternative (Moynihan et al., 2015).

Further research should also consider modifications made to standard essay assessments that are already implemented on some occasions to increase marker reliability. In some cases, the marking procedures can increase

reliability and validity in marking, such as guidance for markers on how to account for their own biases (Alfonso & Flanagan, 2018) or assessing students with many short essays rather than fewer large essays to combat marking fatigue (Jones et al., 2017). Our study did not account for such factors which could reduce our observed effect. Ultimately, if this effect is pervasive, it suggests that essays are more suitable for evaluating smaller groups, where the number of essays being marked cannot induce boredom effects.

Our results suggest a deleterious impact of boredom on marking, with marks becoming lower if general feelings of boredom rise. It stands to reason that similar effects of boredom may be identified in other domains of evaluation, including in academic work. Perhaps peer reviewers may be more inclined to reject contributions when they feel bored, even if this boredom is not necessarily caused by the content of the manuscript under consideration. Likewise, reviewers of grant proposals or conference contributions may be more inclined towards rejection when boredom sets in. This perhaps provocative possibility would require dedicated testing before concluding such undesirable influence of boredom indeed exists.

Conclusion

In conclusion, our study provides evidence that the marking of essays does indeed predict boredom. Following from this, boredom appears associated with more negative evaluations. Our findings suggest that measures should be taken to ensure that the inherent source of boredom in marking essays is sufficiently dealt with. Further research should try to replicate the study with a between-subjects comparison for boredom as well as accounting for similar topics of essays, closer resembling the marking of essays in education.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11031-022-09929-2.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants.

References

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement, 12*(2), 117–128. https://doi.org/10.1177/014662168801200202

Adaval, R. (2003). How good gets better and bad gets worse: Understanding the impact of affect on evaluations of known brands. *Journal of Consumer Research, 30*(3), 352–367. https://doi.org/10.1086/378614

Alfonso, V. C., & Flanagan, D. P. (2018). *Essentials of specific learning disability identification*. Wiley.

Bell, R. C. (1980). Problems in improving the reliability of essay marks. *Assessment in Higher Education., 5*(3), 254–263. https://doi.org/10.1080/0260293800050303

Biggs, J. (1988). Approaches to learning and to essay writing. In *Learning strategies and learning styles* (pp. 185–228). Springer, Boston, MA.

Bird, J. B., Olvet, D. M., Willey, J. M., & Brenner, J. (2019). Patients don't come with multiple choice options: Essay-based assessment in UME. *Medical Education Online, 24*(1), 1–8. https://doi.org/10.1080/10872981.2019.1649959

Bloxham, S. (2009). Marking and moderation in the UK: False assumptions and wasted resources. *Assessment & Evaluation in Higher Education., 34*(2), 209–220. https://doi.org/10.1080/02602930801955978

Bloxham, S., den-Outer, B., Hudson, J., & Price, M. (2016). Let's stop the pretence of consistent marking: Exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education, 41*(3), 466–481. https://doi.org/10.1080/02602938.2015.1024607

Brown, G. (2009). The reliability of essay scores: The necessity of rubrics and moderation. In L. H. Meyer, S. Davidson, H. Anderson, R. Fletcher, P. M. Johnston, & M. Rees (Eds.), *Tertiary assessment and higher education student outcomes: Policy, practice and research* (pp. 40–48). Ako Akotearoa.

Caryl, P. (1999). Psychology examiners re-examined: A 5-year perspective. *Studies in Higher Education., 24*(1), 61–74. https://doi.org/10.1080/03075079912331380148

Cauley, K. M., & McMillan, J. H. (2010). Formative assessment techniques to support student motivation and achievement. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 83*(1), 1–6. https://doi.org/10.1080/00098650903267784

Clore, G. L., & Huntsinger, J. R. (2007). How emotions inform judgement and regulate thought. *Trends in Cognitive Sciences., 11*(9), 393–399. https://doi.org/10.1016/j.tics.2007.08.005

Danckert, J., Hammerschmidt, T., Marty-Dugas, J., & Smilek, D. (2018). Boredom: Under-aroused and restless. *Consciousness and Cognition, 61*, 24–37. https://doi.org/10.1016/j.concog.2018.03.014

De Moira, A. P., Massey, C., Baird, J., & Morrissy, M. (2002). Marking consistency over time. *Research in Education., 67*(1), 79–87. https://doi.org/10.7227/RIE.67.8

Ding, S., & Beaulieu, P. (2011). The role of financial incentives in balanced scorecard-based performance evaluations: Correcting mood congruency biases. *Journal of Accounting Research, 49*(5), 1223–1247. https://doi.org/10.1111/j.1475-679X.2011.00421.x

Eastwood, J. D., Frischen, A., Fenske, M. J., & Smilek, D. (2012). The unengaged mind: Defining boredom in terms of attention. *Perspectives on Psychological Science., 7*(5), 482–495. https://doi.org/10.1177/1745691612456044

Ecclestone, K. (2001). "I know a 2:1 when I see it": Understanding criteria for degree classifications in franchised university programmes. *Journal of Further and Higher Education, 25*(3), 301–313. https://doi.org/10.1080/03098770126527

Feletti, G. I. (1980). Reliability and validity studies on modified essay questions. *Journal of Medical Education, 55*(11), 933–941. https://doi.org/10.1097/00001888-198011000-00006

Forgas, J. P. (1995). Mood and judgment: The affect infusion model (AIM). *Psychological Bulletin., 117*(1), 39–66. https://doi.org/10.1037/0033-2909.117.1.39

Gawron, V. J., French, J., & Funke, D. (2001). An overview of fatigue. In P. A. Hancock & P. A. Desmond (Eds.), *Stress, workload, and fatigue* (pp. 581–595). Lawrence Erlbaum Associates Publishers.

Hack, C. (2013). Using rubrics to improve marking reliability and to clarify good performance. Presented at STEM Conference, Ulster, 2013. University of Ulster. https://www.heacademy.ac.uk/system/files/gen_164_0.pdf

Hayes, A. F. (2009). Beyond baron and kenny: Statistical mediation analysis in the new millennium. *Communication Monographs, 76*(4), 408–420. https://doi.org/10.1080/03637750903310360

Hift, R. J. (2014). Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? *BMC Medical Education, 14*, 249. https://doi.org/10.1186/s12909-014-1249-2

Hunter, A., & Eastwood, J. D. (2018). Does state boredom cause failures of attention? Examining the relations between trait boredom, state boredom, and sustained attention. *Experimental Brain Research, 236*, 2483–2492. https://doi.org/10.1007/s00221-016-4749-7

IBM Corp. Released 2020. *IBM SPSS Statistics for Windows*, Version 27.0. IBM Corp.

Jones, L., Allen, B., Dunn, P., & Brooker, L. (2017). Demystifying the rubric: A five-step pedagogy to improve student understanding and utilisation of marking criteria. *Higher Education Research & Development, 36*(1), 129–142. https://doi.org/10.1080/07294360.2016.1177000

Kayapinar, U. (2014). Measuring essay assessment: Intra-rater and inter-rater reliability. *Eurasian Journal of Educational Research., 14*(57), 113–136. https://doi.org/10.14689/ejer.2014.57.2

Kibble, J. D. (2017). Best practices in summative assessment. *Advances in Physiology Education., 41*(1), 110–119. https://doi.org/10.1152/advan.00116.2016

Klein, J. (2002). The failure of a decision support system: Inconsistency in test grading by teachers. *Teaching and Teacher Education, 18*(8), 1023–1033. https://doi.org/10.1016/S0742-051X(02)00057-4

Klein, J., & El, L. P. (2010). Impairment of teacher efficiency during extended sessions of test correction. *European Journal of Teacher Education, 26*(3), 379–392. https://doi.org/10.1080/02619760320000128201

Malouff, J. M., & Thorsteinsson, E. B. (2016). Bias in grading: A meta-analysis of experimental research findings. *Australian Journal of Education, 60*, 245–256. https://doi.org/10.1177/0004944116664618

Mano, H. (1992). Judgments under distress: Assessing the role of unpleasantness and arousal in judgment formation. *Organizational Behavior and Human Decision Processes, 52*(2), 216–245. https://doi.org/10.1016/0749-5978(92)90036-7

Martin, L. L., & Clore, G. L. (Eds.). (2001). *Theories of mood and cognition: A user's guidebook*. Lawrence Erlbaum Associates Publishers.

McLaughlin, K., Ainslie, M., Coderre, S., Wright, B., & Violato, C. (2009). The effect of differential rater function over time (DRIFT) on objective structured clinical examination ratings. *Medical Education, 43*(10), 989–992. https://doi.org/10.1111/j.1365-2923.2009.03438.x

Meadows, M., & Billington, L. (2005). *A review of the literature on marking reliability*. Report for the National Assessment Agency by AQA Centre for Education Research and Policy. https://research.aqa.org.uk/sites/default/files/pdf_upload/CERP_RP_MM_01052005.pdf

Mizuno, K., Tanaka, M., Yamaguti, K., Kajimoto, O., Kuratsune, H., & Watanabe, Y. (2011). Metal fatigue caused by prolonged cognitive load associated with sympathetic hyperactivity. *Behavioral and Brain Functions, 7*, 17. https://doi.org/10.1186/1744-9081-7-17

Moynihan, A. B., Igou, E. R., Van Tilburg, W. A. P. (2017). Boredom increases impulsiveness: A meaning regulation perspective. *Social Psychology*, *48*(5), 293–309. https://doi.org/10.1027/1864-9335/a000317

Moynihan, A. B., Van Tilburg, W. A. P., Igou, E. R., & Wisman, A. (2015). Eaten up by boredom: Consuming food to escape awareness of the bored self. *Frontiers in Psychology., 6*, 1–10. https://doi.org/10.3389/fpsyg.2015.00369

Newstead, S. (2002). Examining the examiners: Why are we so bad at assessing students? *Psychology Language and Teaching., 2*(2), 70–75. https://doi.org/10.2304/plat.2002.2.2.70

Parmenter, D. A. (2009). Essay versus multiple-choice: Student preferences and the underlying rationale with implications for test construction. *Academy of Educational Leadership Journal, 13*, 57–71.

Pepple, D. J., Young, L. E., & Carroll, R. G. (2010). A comparison of student performance in multiple-choice and long essay questions in the MBBS stage I physiology examination at the University of the West Indies (Mona Campus). *Advances in Physiology Education, 34*(2), 86–89. https://doi.org/10.1152/advan.00087.2009

Prolific [Online Recruitment Platform]. (2019). Prolific.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Ralph, B. C. W., Onderwater, K., Thomson, D. R., & Smilek, D. (2017). Disrupting monotony while increasing demand: Benefits of rest and intervening tasks on vigilance. *Psychological Research Psychologische Forschung, 81*(2), 432–444. https://doi.org/10.1007/s00426-016-0752-7

Reazai, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing., 15*(1), 18–39. https://doi.org/10.1016/j.asw.2010.01.003

Samuels, L. B., & Coffinberger, R. L. (2004). Balancing the needs to assess depth and breadth of knowledge: Does essay choice provide a solution. *Journal of Legal Studies Education, 22*(2), 103–122. https://doi.org/10.1111/j.1744-1722.2005.00014.x

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing., 25*(4), 465–493. https://doi.org/10.1177/0265532208094273

Schwarz, N. (2012). Feelings-as-information theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (pp. 289–308). Sage Publications Ltd. https://doi.org/10.4135/9781446249215.n15

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of

affective states. *Journal of Personality and Social Psychology, 45*(3), 513–523. https://doi.org/10.1037/0022-3514.45.3.513

Shackleton, V. J. (1981). Boredom and repetitive work: A review. *Personnel Review., 10*(4), 30–36. https://doi.org/10.1108/eb055445

Slomp, D. H. (2012). Challenges in assessing the development of writing ability: Theories, constructs and methods. *Assessing Writing., 17*(2), 81–91. https://doi.org/10.1016/j.asw.2012.02.001

Smith, C. A., & Ellsworth, P. C. (1987). Patterns of appraisal and emotion related to taking an exam. *Journal of Personality and Social Psychology, 52*(3), 475–488. https://doi.org/10.1037//0022-3514.52.3.475

Smith, E., & Coombe, K. (2006). Quality and qualms in the marking of university assignments by sessional staff: An exploratory study. *Higher Education, 51*, 45–69. https://doi.org/10.1007/s10734-004-6376-7

Tempelaar, D. T., Heck, A., Cuypers, H., van der Kooij, H., & van de Vrie, E. (2013). *Formative assessment and learning analytics.* Paper presented at the Third International Conference on Learning Analytics and Knowledge, Leuven, Belgium. https://doi.org/10.1145/2460296.2460337

Thompson, C., Fransen, J., Beavan, A., Skorski, S., Coutts, A., & Meyer, T. (2020). Understanding the influence of a cognitively demanding task on motor response times and subjective mental fatigue/boredom. *Brazilian Journal of Motor Behavior, 14*, 33–41. https://doi.org/10.20338/bjmb.v14i01.167

Van Der Vleuten, C. P. (2016). Revisiting 'Assessing professional competence: From methods to programmes.' *Medical Education., 50*(9), 885–888. https://doi.org/10.1111/medu.12632

Van Der Vleuten, C. P., & Schuwirth, L. W. (2005). Assessing professional competence: From methods to programmes. *Medical Education., 39*(3), 309–317. https://doi.org/10.1111/j.1365-2929.2005.02094.x

Van Tilburg, W. A. P., & Igou, E. R. (2012). On boredom: Lack of challenge and meaning as distinct boredom experiences. *Motivation and Emotion, 36*(2), 181–194. https://doi.org/10.1007/s11031-011-9234-9

Van Tilburg, W. A. P., & Igou, E. R. (2016). Going to political extremes in response to boredom. *European Journal of Social Psychology, 46*(6), 687–699. https://doi.org/10.1002/ejsp.2205

Van Tilburg, W. A. P., & Igou, E. R. (2017). Boredom begs to differ: Differentiation from other negative emotions. *Emotion, 17*(2), 309–322. https://doi.org/10.1037/emo0000233

Van Tilburg, W. A. P., Igou, E. R., & Sedikides, C. (2013). In search of meaningfulness: Nostalgia as an antidote to boredom. *Emotion, 13*(3), 450–461. https://doi.org/10.1037/a0030442

Westgate, E. C., & Wilson, T. D. (2018). Boring thoughts and bored minds: The MAC model of boredom and cognitive engagement. *Psychological Review, 125*(5), 689–713. https://doi.org/10.1037/rev0000097

Williams, R., Sanford, J., Stratford, P. W., & Newman, A. (1991). Grading written essays: A reliability study. *Physical Therapy., 71*(9), 679–686. https://doi.org/10.1093/ptj/71.9.679

Wyatt-Smith, C., & Klenowski, V. (2013). *Assessments for education: Standards, judgement and moderation*. SAGE Publications Ltd. https://doi.org/10.4135/9781526401878

Wyer, R. S., Clore, G. L., & Isbell, L. M. (1999). Affect and information processing. *Advances in Experimental Social Psychology, 31*, 1–77. https://doi.org/10.1016/S0065-2601(08)60271-3

Yang, J. C. (1987). Reliability of grading essay papers in a baccalaureate nursing programme. *Nurse Education Today., 7*(3), 120–125. https://doi.org/10.1016/0260-6917(87)90099-2