Putwain, DW, Nicholson, L and Kutuk, G

Warning Students of the Consequences of Examination Failure: An Effective Strategy for Promoting Student Engagement?

Warning Students of the Consequences of Examination Failure: An Effective Strategy for Promoting Student Engagement?

David W. Putwain[1], Laura J. Nicholson[2], and Gulsah Kutuk[3]
[1]School of Education, Liverpool John Moores University, Liverpool, UK.
[2]Department of Psychology, Edge Hill University, Ormskirk, UK.
[3]School of Education and Lifelong Learning, University of East Anglia, Norwich, UK.

Author Note

David W. Putwain https://orcid.org/0000-0001-5196-4270
Laura J. Nicholson https://orcid.org/0000-0001-9255-9650
Gulsah Kutuk https://orcid.org/0000-0003-2921-5358

We have no conflicts of interest to disclose. The research was funded by a Faculty Research Development Award made to the second author. Data, materials, and analysis are, are available by emailing the corresponding author. This study was not preregistered.

Address for correspondence: Professor David W. Putwain, Liverpool John Moores University, Maryland St, Liverpool, L1 9DE, UK. Email d.w.putwain@ljmu.ac.uk

Abstract

In the context of high-stakes qualifications, teachers may warn students of the negative consequences of failure as a tactic designed to increase engagement and ultimately achievement. Previous studies have shown that these types of messages, namely fear appeals, are indirectly related to engagement and achievement in different ways depending on how they are evaluated by the student. When fear appeals are evaluated as a challenge, they are positively related to engagement and achievement. When evaluated as a threat, fear appeals are negatively related to engagement and achievement. In the present study, we offer a robust test of these relations in a multi-level model that controls for autoregressive and concurrent relations in the domain of mathematics. Self-reported data were collected from 1,530 participants, aged 14-16 years, at two time points over the final two years of secondary education. These data were linked to prior and subsequent achievement. Results showed that students who attended to fear appeals and evaluated them as a challenge showed higher subsequent engagement, and students who showed higher engagement showed higher achievement. Accordingly, it may be beneficial to identify those students likely to evaluate fear appeals as a threat and intervene in order to enhance the likelihood of a challenge evaluation (e.g., building confidence through strategy focused feedback and strengthening beliefs in the value of effort). Given the difficulties associated with teachers judging students' motivation and emotion as private experiences, methods to access student voice should be considered.

Keywords: Fear appeals; challenge; threat; mathematics; achievement

Educational Impact and Implications Statement

Teachers may communicate messages with a motivational impact, including negative messages (i.e., fear appeals), in the context of high-stakes qualifications. It is, therefore, useful to know and understand which messages may be more or less effective and under which conditions. Our study showed that if evaluated as a challenge (as opposed to threat), fear appeals could be an efficacious means to promoting students' engagement. However, given the difficulty of identifying and targeting individual students' responses to these messages, fear appeals should be used cautiously by potentially implementing a personalized approach.

Warning Students of the Consequences of Examination Failure: An Effective Strategy for Promoting Student Engagement?

In many areas of life, persons encounter messages that intentionally, or otherwise, indicate the possible outcomes of different courses of action. Some messages provide information of the courses of actions expected to result in desired outcomes. Other messages focus on those actions that could result in undesirable consequences. Such messages are often encountered in the public health domain, for instance, to promote behaviors likely to avoid ill-health or injury (Ruiter et al., 2014). At the time of writing, messages regarding behaviors believed to stop infection and spread of the COVID-19 disease (e.g., social distancing and mask wearing) are ubiquitous (Breakwell et al., 2021; Vally, 2020). However, such messages are also frequently used and found in other areas of life (e.g., financial planning, cyber information security, climate change, and so on). Messages can be purely informational, but take on a particular relevance for the recipient when consequences are central to that person's goals (Lazarus, 2001).

Classrooms are no exception. Teachers may use messages routinely within their instructional repertoire to suggest to students advantageous approaches to classroom tasks or activities, homework assignments, examination preparation, and when providing solution focused feedback on student work (e.g., Howe et al., 2019; Mercer & Littleton, 2007). Messages may also be used as a method of maintaining classroom discipline to remind students of the consequences of disruptive behavior (e.g., Kearney et al., 1985; Richmond, & McCroskey, 1992). In the present study, we focus on messages used by teachers in the run-up to high-stakes secondary school-exit qualifications (see Banks & Smyth, 2015; Sprinkle et al., 2006).

In England, where the present study was based, students take national standardized secondary school exit-examinations aged 16 years[1] leading to qualifications in the General Certificate of Secondary Education (GCSE). GCSE qualifications have profound and far-reaching consequences for students. Access to upper/post-secondary education (school or college-based courses in academic, vocational, technical, education or work-based apprenticeships) is dependent on a profile of GCSE grades (Long & Bolton, 2017; Shackleton, 2014). Furthermore, entry requirements for all occupations (other than routine and manual labor), require minimum pass grades in English and mathematics, often in conjunction with minimum grade requirements in other subjects (Maguire, 2010; Roberts, 2004).

Students in England who do not achieve minimum pass-grades in English language and mathematics must continue to re-take GCSE examinations until the age of 18 years, at which point they can legally leave education, or until they pass, at which point they can enroll for a course in upper/post-secondary education. In short, the stakes of the GCSE are high for students and it is not difficult to see how messages about the GCSE outcomes would have high personal relevance for most. There are similar high-stakes secondary school-exit examinations taken in other parts of the world including the United States and Western Europe (e.g., Bishop & Mane, 2001; Carnoy, 2005; Heath et al., 2008).

GCSEs may be high-stakes for teachers as well as for students. Secondary schools are inspected once every three years in England and judgments traditionally placed a strong emphasis on GCSE achievement as a measure of accountability (see Roberts & Abreu, 2016; Perryman et al., 2011)[2]. Furthermore, schools are ranked within localities into league tables, to use a sporting analogy, based on their GCSE achievement (Perryman, 2006). Given these

[1] GCSE examinations are also taken in Wales and Northern Ireland. Examinations were replaced during the COVID-19 pandemic (2020 and 2021) with teacher estimated grades.
[2] A new inspection framework was introduced in 2019 with greater emphasis (compared to previous) on wellbeing.

high-stakes, it is perhaps not surprising that teachers frequently communicate messages about the importance of success and/or avoiding failure to their students (Putwain & Roberts, 2012; Putwain & von der Emsbe, 2018).

In the present study, we focus specifically on those messages that highlight the negative consequences of failure, or the importance of avoiding failure. For brevity, and to be consistent with the extant literature, we refer to these as fear appeals. Studies have shown that fear appeals can be evaluated in different ways by students; as a challenge if the message is judged as personally meaningful and the student believes they can avoid failure, and as a threat if judged as meaningful and the student believes failure is likely (Symes & Putwain, 2016). It might be expected that the impact of fear appeals on downstream achievement-related behavior and examination grades would depend on whether the message was evaluated as a challenge or as a threat. Given the central role of the teacher as an agent of learning in the classroom (e.g., Hattie, 2009; Wentzel et al., 2009), it is critical for teachers to recognize whether the language used with students is likely to help or hinder student engagement and achievement. Such awareness could be easily disseminated to teachers and implemented to good effect in a cost-effective manner.

Teacher Motivational Messages

In the program of study leading up to high-stakes qualifications (such as the GCSE), teachers and school leaders communicate information about those examinations to students. Some of this information will be purely administrative (e.g., when coursework must be submitted, the time and date of the examination, which curricular topics are examined in which paper, and so on) and some of this information may relate to study skills and effective examination preparation. However, teachers can also, intentionally or otherwise (see Flitcroft et al., 2018), communicate motivational information about forthcoming examinations.

Motivational messages have two dimensions corresponding to the message-frame and the type(s) of motivation the message appeals to (Putwain & Woods, 2016). Frame refers to whether the outcome in a message is a gain or loss (Rothman & Salovey, 1997). Gain-focused messages (success appeals) focus on the consequences of success and loss-focused messages (more commonly referred to as fear appeals) focus on failure or avoiding failure. Potentially, different types of motivations could be highlighted in a message. In the present study, we chose to focus on value, as the type of motivation appealed to, as reflecting the way that teachers in England commonly communicate the importance of GCSEs to students. For instance, teachers may highlight to students how GCSE grades can influence future life trajectory (e.g., earning potential), and educational and occupational aspirations (e.g., Flitcroft et al., 2018; Putwain, 2009).

Accordingly, we draw on motivational theories that emphasize utility or extrinsic value, namely Expectancy Value Theory (EVT: Wigfield et al., 2016) and Control-Value Theory (CVT: Pekrun, 2018). Utility and extrinsic value are conceptualized in EVT, and CVT respectively, as subjectively valuing a task or activity when it is appraised as being instrumental to a student's short or long-term goals. Motivational messages could, however, also be approached from a self-determination theory perspective. Teachers could, for example, appeal to less (working hard for a reward or to feel proud) or more (working hard to prepare for higher-level study or master a topic) self-determined forms of motivation (e.g., Monagas et al., 2021).

The majority of studies examining teacher motivational messages in the context of high-stakes qualifications have focused on loss-focused messages (for a review see Putwain et al., 2021). Furthermore, the few studies that have compared gain- with loss-focused messages found that participants evaluate them in very similar ways (e.g., Putwain & Symes, 2014, 2016). To use a common idiom, they are two sides of the same coin; that is, emphasizing the possible benefits of success defacto implies the likely drawbacks of failure and vice versa. At present, evidence is lacking to show whether gain- and loss-focused messages used in the context of high-stakes qualifications, and their outcomes, lead to differential outcomes. In order to avoid using a duplicate set of similarly worded items (one set each for gain- and loss-focused messages) thereby increasing participant burden and potentially resulting in neglectful answers, we chose in the present study to focus on just one message, namely loss-focused value messages.

For brevity, and to be consistent with the extant literature, we refer to loss-focused value messages as fear appeals. Fear appeals, used in the context of high-stakes qualifications, are defined as messages that highlight the negative consequences of failure along with actions that are likely to increase or reduce failure (Putwain & Symes, 2014). These messages are used relatively frequently by teachers in the belief that students will respond positively by increasing their motivation, engagement, and effort (Putwain & Roberts, 2012; Putwain & von der Embse, 2018).

Student Evaluation of Teacher Fear Appeals

In keeping with appraisal frameworks, notably, EVT (Wigfield & Eccles, 2020; Wigfield et al., 2016), CVT (Pekrun, 2018; Pekrun & Perry, 2014), and the Cognitive Theory of Stress and Coping (CTSC: Folkman & Lazarus, 1985; Lazarus & Folkman, 1984), we propose that teacher fear appeals will not directly link to student motivation, engagement, and achievement, but will depend on how they are evaluated by students (also see Kuppens & Tong, 2010; Travis et al., 2020).

We expect readers will already be familiar with EVT (e.g., Wigfield & Eccles, 2020), CVT (e.g., Pekrun, 2018), and CTSC (e.g., Lazarus & Folkman, 1984), and so will not expand on these theories here other than to briefly state the central tenets. In EVT, expectancy of success combines with subjective task value to influence educational choice, engagement, and achievement. In CVT, control over achievement activities and outcomes combines with subjective task value to influence emotions and subsequent motivation, information processing, and achievement. What we can surmise from EVT and CVT is that stronger beliefs in one's competence (which underlie expectancy and control appraisals) are associated with positive achievement emotions, behaviors, and educational outcomes.

In CTSC, events or situations, in a primary appraisal, are judged to be irrelevant, benign (a likely positive outcome), or stressful (the potential for harm or growth). Options and resources for coping with the stressful situation are judged in a secondary appraisal. Challenge arises when coping resources are judged to outweigh task demands resulting in energized motivation, positive emotions, and improved performance. Threat arises when coping resources cannot meet task demands resulting in impaired information processing, negative emotions, and reduced performance. It remains unclear at present the extent to which challenge and threat remain opposite states (as originally conceptualized in CTSC) or whether it is possible for them to co-occur (e.g., Nicholson et al., 2019; Uphill et al., 2019).

We can deduce from EVT, CVT, and CTSC, that messages regarding high-stakes qualifications are likely judged in terms of their personal importance; that is, whether the consequences of success or failure will impact one's aspirations (i.e., utility value). In addition, coping resources, such as the perceived competence of the person (e.g., in a subject domain or study skills), will determine whether the person responds to the message with adaptive behaviors (e.g., increased effort in class) or not.

Accordingly, we propose that fear appeals could be evaluated, to use the parlance of CTSC, as a challenge or threat. We prefer the term evaluation over appraisal to capture not only the role of beliefs in how fear appeals are understood and responded to, but emotions and behavioral intentions. As noted in CTSC, evaluations that lead to challenge and threat states may be partly or wholly automatized. Under such circumstances, emotions and behavioral intentions, as well as beliefs, may be used by the person to understand their reactions to fear appeals (see Schwarz & Clore, 1983; Tiedens & Linton, 2001).

Experimental (Putwain & Symes, 2014, 2016) and naturalistic studies (Putwain & Remedios, 2014; Putwain, Remedios, et al., 2016; Symes & Putwain, 2016) using samples of secondary school students have shown that challenge follows from higher self-efficacy (an indicator of domain-specific competence beliefs) and higher utility/attainment value; threat arose from lower self-efficacy and higher utility/attainment value. The fear appeal message acts as a trigger for evaluative processes resulting in stronger and self-reinforcing challenge and threat evaluations when fear appeals are made more frequently. Accordingly, message frequency is associated with greater challenge and threat evaluations (Putwain, Symes, et al., 2016; Symes et al., 2015).

Teacher Fear Appeals, their Evaluation, and Subsequent Motivation, Emotion, and Behavior

Theoretically speaking, we would expect fear appeals to result in positive appetitive motivations, achievement emotions (e.g., hope), and achievement behaviors (e.g., greater engagement), if evaluated as a challenge (Putwain et al., 2021). Conversely, if evaluated as a threat, fear appeals would result in negative avoidant motivations, achievement emotions (e.g., anxiety), and achievement behaviors (e.g., withdrawal of effort). That is, relations between fear appeals and subsequent achievement-related motivation, emotion, and behavior, would be indirect and mediated by fear appeal evaluations. Ultimately, fear appeals would also relate to greater educational achievement if evaluated as a challenge and worse educational achievement if evaluated as a threat due to the impact of positive/negative achievement-related motivation, emotion, and behavior. That is, relations between fear appeals and achievement would be serially mediated by evaluations and subsequent motivation, emotion, and behavior.

Relatively few studies have examined relations from fear appeals, and their evaluations, to subsequent engagement and achievement. The few exceptions that have done so support the proposition that challenge evaluations are positively, and threat evaluation, negatively related to engagement and achievement (Nicholson & Putwain, 2020; Putwain et al., 2017; Putwain, Nicholson, et al., 2016; Putwain & Symes, 2011). The aforementioned studies are, however, characterized by methodological and analytic weaknesses. Three of the studies did not include a temporal separation between challenge/threat evaluation and engagement/motivation (Putwain et al., 2017; Putwain, Nicholson, et al., 2016; Putwain & Symes, 2011) and the fourth study included evaluations of fear appeal but not frequency (Nicholson & Putwain, 2020).

In the present study, we set out to address these weaknesses in a sample of secondary school students preparing for their high-stakes GCSE qualifications using a robust multi-level analysis. Over two waves, we examined how fear appeals and their evaluations were related to subsequent achievement. That is, we examine how fear appeals at the first wave (T_1) were related to their evaluations at the second wave (T_2) of data collection, and how T_1 fear appeal evaluations were related to T_2 behavioral engagement, while controlling for un-lagged and auto-lagged relations. In addition, we examined relations to subsequent achievement, allowing for tests of whether fear appeals were indirectly related to achievement, mediated by fear appeal evaluations and behavioral engagement. By adopting a multi-level approach, we also address a question not previously considered, namely whether it is not only fear appeals per se that might influence behavioral engagement and achievement, but the extent to which students attend more closely to these fear appeals. We address this point more fully next.

The Multi-level Question

Fear appeals made by a teacher to a whole class are, defacto, a classroom climate (L2) variable. Classroom climate variables can be measured from different informants (e.g., teacher self-report or student observer reports). Each perspective offers a unique insight into the classroom environment and is associated with different biases and accuracy (e.g., Kunter & Baumert, 2006; Paulus & Vazire, 2010). In the present study, we used student reports. The advantage of this approach is that it relies on multiple observers and, notwithstanding issues such as teacher rapport, is potentially more accurate than that of a single report from a teacher.

It is therefore necessary to aggregate student reports for each class and use the mean score for that class as the classroom climate (L2) variable after they have been grand mean-centered (see Hox, 2010). A further advantage of this approach is that the reliability of the aggregated variable can be established using the ICC_2 statistic (see Lüdtke et al., 2009). Previous studies using aggregated student reports of teacher fear appeals have reported high ICC_2 values (ICC_2s = .84 and .91, in Putwain, Remedios, et al., 2016, and Putwain, Symes, et al., 2017, respectively).

We theorized that any relations between fear appeals and subsequent variables, such as engagement and achievement, would be mediated by fear appeal evaluation. The evaluation of a fear appeal by a student, and behavioral engagement, are L1 variables. However, any potential mediation of a classroom climate (L2) variable must also necessarily be an L2 variable. That is, because the mean aggregated fear appeal for a particular class is constant, there is no variation with which to model relations with L1 fear appeal evaluations (Hofmann, 2002). The solution is to test mediational relations within a ML-SEM that analyses the between- and within-level variance within observed variables separately (see Preacher et al., 2010). Of course, this analytic-conceptual rationale depends on sufficient between-level variance in fear appeal evaluation and engagement to warrant modelling multi-level relations.

In order to examine multi-level indirect relations from fear appeals to engagement and GCSE grade, scores for fear appeal evaluations and behavioral engagement were grand mean-centered and aggregated within classes. Class-aggregated scores for L1 constructs, where the student is the referent (like fear appeal evaluation and engagement) differ from climate variables where the referent is a classroom phenomenon, in our case, teacher behavior (see Marsh et al., 2009, 2012; Morin et al., 2014). It is possible that L1 variables may result in different effects when class-aggregated (e.g., assimilation and contrast effects as shown in the big-fish-little-pond effect; Preckel & Brüll, 2010). To establish a 'true' contextual effect, it is necessary to estimate the effect of the L2 variable beyond that of the L1 variable (Marsh et al., 2009, 2012; Morin et al., 2014). Accordingly, in the present study, new coefficients for contextual effects were estimated by subtracting L1 from L2 effects in an additional, supplementary, analysis.

In the within-level portion of the ML-SEM, relations can still be modeled between class mean-centered fear appeal evaluations, engagement, and achievement; we just cannot make the prior link to teacher fear appeals. Within-level student reports of fear appeals (i.e., whether the student reports their teacher using a fear appeal more or less frequently than the class average), however, are not necessarily completely redundant. They do, nevertheless, require careful consideration as to what exactly they measure or indicate (Marsh et al., 2009, 2012). Putwain and Best (2012) showed that when exposed to fear appeals about a forthcoming test in an experimental manipulation, primary school children showing high test anxiety (measured one month previous) reported their teacher to use more frequent fear appeals than their low test-anxious peers. This is a likely consequence of test anxiety increasing vigilance for environmental threat cues (Putwain et al., 2020). Accordingly, we propose that class mean-centered within-level student reports of fear appeals are an indicator of the extent to which a student attends to the teacher message. In the absence of a prior measure to account for potential bias in attention focus, we do not attach a specific reason for within-level variance in student reports of teacher fear appeals; a student could attend to a fear appeal due to high motivation as well as fear.

To summarize our approach to the multi-level issue, we address the question of how fear appeals are indirectly related to engagement, and achievement, using a ML-SEM. In the between-level portion of the model, we examine relations between class-aggregated fear appeals, fear appeal evaluations, engagement, and achievement. In the within-level portion of the model, we examine relations between class mean-centered fear appeals, fear appeal evaluations, engagement, and achievement.

Aims of the Present Study

The aim of the present study was to examine indirect relations from fear appeals to achievement mediated by fear appeal evaluations and behavioral engagement using a robust design and analytic approach. In a two-wave design, predictor, mediator, and outcome variables are all measured at both time points (Cole & Maxwell, 2003; Little et al., 2007). Relations between predictor and mediator variable (path *a*), and between mediator and outcome variables (path *b*), are established using the same temporal lag from Time 1 (T_1) to Time 2 (T_2). Indirect relations are established from the product of paths *a* and *b*. Including concurrent correlations between predictor, mediator, and outcome variables, adds an additional level of analytic robustness by controlling for prior relations between predictor, mediator, and outcome variables.

The present study, therefore, allowed a temporal separation between T_1 fear appeals and T_2 evaluations, and from T_1 evaluations and T_2 behavioral engagement. These data were linked to prior and subsequent achievement. Using a ML-SEM, we were able to assess relations at the classroom level (i.e., the between portion of the model to examine relations using class-aggregated variables), as well as at the student level (i.e., the within portion of the model to examine relations using class mean-centered variables). Due to the likely influence of socio-demographic variables, we controlled for gender, age, and the eligibility of participants for free school meals (FSM) as a proxy for a student coming from a low-income household. The following three hypotheses were tested at the within (L1) level using class mean-centered scores and the between (L2) level using class-aggregated scores (see Figure 1):

Hypothesis 1: Fear appeals will be positively related to challenge and threat evaluation.

Hypothesis 2: Challenge evaluation will be positively, and threat evaluation negatively, related to behavioral engagement.

Hypothesis 3: Behavioral engagement will be positively related to subsequent mathematics grade.

Three indirect paths were estimated at the within- (L1) and between-level (L2) portions of the model. First, indirect relations were tested using the direct paths from T_1 fear appeals to T_2 challenge and threat evaluation (labeled paths *a*1 and *a*2 respectively in Figure 1), and from T_1 challenge and threat evaluation to T_2 behavioral engagement (labeled paths *b*1 and *b*2 respectively in Figure 1). Second, we tested indirect paths from T_1 challenge/threat evaluation to GCSE grade, via T_2 behavioral engagement (the path from behavioral engagement to GCSE grade is labeled *b*3 in Figure 1). Third, we tested indirect paths from T_1 fear appeals to GCSE grade, mediated by T_2 challenge/threat evaluation and T_2 behavioral engagement.

As the constructs under investigation are domain-specific (e.g., Green et al., 2007; Hamre et al., 2014), in order to maintain specificity matching (see Swan et al., 2007), we focused on a single subject, namely mathematics. Although many students experience high levels of mathematics anxiety (Maloney, 2016), learning-related anxieties can also occur in other subjects including foreign language learning (Teimouri et al., 2019), science (Sinatra et al., 2016), and those that require reading in one's native language (e.g., Piccolo et al., 2017). Thus, we would not necessarily expect a differential response to fear appeals in mathematics lessons than other subjects based purely on the nature of the subject.

The stakes associated with mathematics may, however, exert an influence. As noted above, a minimum pass grade in GCSE mathematics, along with English, is a pre-requisite for many occupations beyond routine and manual (Maguire, 2010), and required for post-secondary training (Shackleton, 2014). Indeed this point is likely to form the basis of fear appeals that emphasize the value of mathematics (Putwain et al., 2021). We might, therefore, expect fear appeals to be more impactful in mathematics, as well as English, making these subjects particularly important to study fear appeals in relation to. The impact of fear appeals in other subjects may depend more on an alignment with an individual student's interests and aspirations.

Method

Participants

The participants in this study were a convenience sample of 1,530 students from 14 secondary schools located in the North West of England. The study spanned two school years. At the outset, participants were in Year 10 (the penultimate year of secondary school) with a mean age of 14.6 years (*SD* = .49) and at the end of the study in Year 11 (the final year of secondary school) with a mean age of 15.6 years (*SD* = .50). Gender was relatively evenly matched (T_1: 50.7% female). The proportion of students from Black and Minority Ethnic backgrounds (T_1: 17.8%) was lower than the average of 31% for all English secondary schools in 2019 (Department for Education, 2019). We also collected data about the number of pupils eligible for FSM at the outset of the study. There were 299 students eligible at T_1 (21.1% after accounting for 105 missing responses). This was a higher proportion than the average of 12.4% for all English secondary schools in 2019 (Department for Education, 2019).

At T_1, students were clustered into 105 different mathematics classes with a mean of 14.5 participants per class and at T_2, 94 different mathematics classes with a mean of 11 participants per class. In 2019, there was a mean of 21.7 students per class in English secondary schools (Department for Education, 2019). In common with many longitudinal studies (Coertjens et al., 2017), there was substantial attrition from T_1 to T_2 (32.5%) resulting from the combination of students being absent from school or choosing not to participate. Having established that data were Missing at Random (MAR), Full Information Maximum Likelihood (FIML) was used in subsequent analyses to manage missing data. See Supplementary Materials for additional details regarding participant characteristics across the two waves of data collection and missing data analyses. Classroom composition was relatively stable across T_1 and T_2; of the 1,032 students reporting data at both waves, 71 changed class (6.9%).

Measures

Frequency and Evaluation of Fear Appeals

The frequency and evaluation of fear appeals was measured using nine items from the *Teachers Use of Fear Appeals Questionnaire* (Putwain et al., 2019). Three items measured the frequency of fear appeals (e.g., "How often does your maths teacher tell your class that unless you work hard you will fail your maths GCSE?" followed by "If your maths teacher says this, do you…")[3]. Challenge and threat evaluation were measured by three items each. The three challenge evaluation items referred to interpreting fear appeals as inspiring hard work, effort, and hope (e.g., "feel inspired to work hard to pass GCSE maths"). The three threat evaluation items referred to interpreting fear appeals as worrying, indicating likely

[3] In everyday parlance in the UK, mathematics is referred to as 'maths'.

failure, and struggling to pass (e.g., “feel worried by the possibility of failing GCSE maths”). Participants responded to items on a five-point scale (1 = *never*, 2 = *occasionally*, 3 = *sometimes*, 4 = *quite a lot of the time*, and 5 = *most of the time*). Previous research has supported the factorial validity and internal consistency of data collected using this measure (e.g., Putwain et al., 2019) and earlier iterations (Putwain & Symes, 2014). In the present study, the internal consistency was strong (McDonald’s ω ≥.80; see Table S2).

Behavioral Engagement

Behavioral engagement was measured using three items from the *Engagement vs. Dissatisfaction with Learning Questionnaire* (Skinner et al., 2009). Participants responded to items, adapted to specifically refer to GCSE mathematics (e.g., ‘I participate in the activities and tasks in my GCSE maths class’), on a five-point scale (1 = *strongly disagree* to 5 = *strongly agree*). The original behavioral engagement scale consisted of five items. Preliminary analyses, however, indicated correlated residual variances between two pairs of items (“I try hard to do well in my GCSE maths class” with “In my GCSE maths class, I work as hard as I can” and “I pay attention in my GCSE maths class” with “When I’m in my GCSE maths class, I listen very carefully”). Post-hoc addition of correlated residual variance can be justified when resulting from method effects such as the wording of items (see Cole et al., 2007), the likely cause of the correlated residual variance here. However, we opted to drop one item from each of the pairs of items showing correlated residual variance to avoid additional parameters in an already complex model. Data collected using the five-item and reduced three-item versions of the behavioral engagement scale have shown construct validity, predictive validity, and internal consistency, in previous studies (e.g., Skinner et al., 2008, 2009; Skinner & Chi, 2012). In the present study, the internal consistency was strong (McDonald’s ω ≥.87; see Table S2).

Mathematics Achievement

GCSE Mathematics Achievement. GCSE mathematics was graded on a 9-point numeric scale. Grade 9 is the highest and grade 1 the lowest; grade 4 is considered a pass. Grades are usually awarded on the basis of examination marks. Grade boundaries are set using a combination of criterion- and norm-referenced approaches to keep the distribution of grades broadly similar from one year to the next (Office of Qualifications and Examinations Regulation, 2016).

Following school closures in England on 18th March 2020 due to the COVID-19 pandemic, all GCSE examinations scheduled for May and June 2020 were cancelled. GCSE grades were awarded on the basis of teacher predictions using internal quality procedures within schools to ensure consistency between different teachers and moderated at the school level by the examination regulator (Office of Qualifications and Examinations Regulation). Accordingly, we were unable to make use of grades from GCSE examinations as planned, but used the officially awarded teacher estimated grades as a proxy.

The education system in England (as well as the rest of the UK) is relatively unique in the routine use of teacher estimated grades for official purposes to determine offers for university study in advance of formal examination grades (Anders et al., 2020). The standardization of grading processes within and between schools will likely be harder to achieve than if examinations were marked and graded by an external awarding body (as usually happens). We are confident, however, that the use of grade criteria in GCSE mathematics, along with the prior experience of teachers in predicting grades, and the quality procedures used within and between schools, add a level of robustness and reliability to the grading process.

Year 9 Mathematics Achievement. Year 9 mathematics achievement was taken from an end-of-year examination, teacher assessment based grades, or a combination of the two. Whichever approach was adopted, achievement was assessed using GCSE criteria, and awarded a grade from 1 to 9.

Demographic Variables

Gender (0 = male, 1 = female), age, and FSM as a proxy for low income (0 = not eligible, 1 = in receipt of FSM), were included as covariates in the analyses.

Procedure

Schools were recruited from a pool of secondary schools in partnership with the second author's institution or who had existing research relationships with the authorship team. Letters asking for expressions of interest in the project were sent in October/November 2018 to the head teachers. In participating schools, the mathematics GCSE program of study was studied over Years 9 to 11 (the final three years of secondary education), culminating in national examinations scheduled over May and June when students were in Year 11. Accordingly, we scheduled self-reported data collection over two waves. T_1 data collection was taken midway through Year 10 (the penultimate year of secondary education) in March/April 2019 and T_2 data collection when students were in Year 11 (the final year of secondary education) in October 2019. Mathematics achievement from Year 9 examination grades (June 2018) and GCSE grades (awarded in August 2020) were taken from official school records.

Data were collected in school in a period of the timetable used for administration by the regular class teacher following a standardized script that emphasised the voluntary nature of participation and that the data would remain confidential and anonymous. Schools were offered the option of electronic or paper and pencil data collection. For electronic data collection, the questionnaire was hosted on a survey data collection site and participants provided with the uniform resource locator to allow completion on a PC, laptop, tablet, or smartphone. Participants' questionnaire responses at the two time points were matched using a memorable code made up by the participant and linked to mathematics achievement data using their unique candidate number[4]. Written institutional consent was provided by the head teachers of participating schools. Individual consent was provided by participants that was either written (for paper and pencil questionnaires) or via the click of a button to indicate consent (for the electronic data collection). Ethical approval for this study was provided by the Faculty Research Ethics Committee at the second author's institution (FOE18-LN01).

Regarding the methods of assessment that teachers used to estimate GCSE grades, the guidance provided by the examinations' regulator allowed teachers and schools to exercise autonomy (Office of Qualifications and Examinations Regulation, 2020). Assessments could include homework assignments and mock exams completed virtually during the period March to May 2020, before estimated grades were submitted in June 2020, and existing records of student performance over the course of study. As T_2, self-report data were collected near the beginning of the school year, the greater part of existing Year 11 performance data would have come from after this point. Indeed, a common practice in many English schools is to hold mock examinations under standardized GCSE test conditions in December or January of Year 11. It is likely the existing student performance used by teachers would have drawn on scores from these examinations, taken after T_2 self-report measures.

Analytic Plan Summary

Analyses proceeded in three stages. First, in a series of preliminary analysis, we examined assumptions regarding missing data, estimated descriptive statistics, and checked basic pre-requisites for multi-level modelling (the variance components of variables and the consistency of fear appeal frequency as a class climate variable). Second, we estimated a

[4] These are unique identifier numbers provided to schools and students in Years 10 and 11 by the GCSE awarding bodies (there are three such bodies approved by the Department for Education in England).

measurement model and latent bivariate correlations using a multi-level confirmatory factor analysis (ML-CFA) and tested longitudinal measurement invariance for fear appeal frequency and evaluation, and behavioral engagement, across T_1 and T_2 waves of data collection, in a series of ML-CFAs. Third, and finally, a ML-SEM was used to test hypotheses.

ML-CFAs and ML-SEMs were estimated using the M*plus* v.8 software (Muthén & Muthén, 2017). Goodness-of-fit was assessed using the root mean error of approximation (RMSEA), standardized root mean residual (SRMR), comparative fit index (CFI), and Tucker-Lewis index (TLI). Based on simulation studies, Hu and Bentler (1999) recommend RMSEA ≈ .06, SRMR ≈ .08, and CFI/TLI ≈ .95, for a good model fit. Pagett and Morgan (2021) also cautiously recommend SRMR indices of ≈.08 for multi-level models under optimal conditions (i.e., number of L2 units >100); under less optimal conditions, $SRMR_B$ >.08 may not necessarily indicate model misspecification when the number of L2 units is <100. A more detailed description of the analytic plan can be found in the Supplementary Materials.

Results

Descriptive Statistics

Descriptive statistics are reported in Table 1. The skewness and kurtosis of variables was within ±1,with the exception of T_2 behavioral engagement, and internal consistency was good (McDonald's $\omega \geq .79$). The proportion of variance attributable to the class level warranted a multi-level approach to modelling data (ICC_1s = .22 and .30 for fear appeal frequency, .28 and .29 for mathematics grades, and .01 to .14 for fear appeal evaluations). ICC_2 statistics showed L2 fear appeal frequency, built from aggregated student responses, were reliable (.80 and .86 for T_1 and T_2, respectively). A more detailed presentation of the descriptive statistics can be found in the Supplementary Materials.

Testing a Measurement Model, Correlations, and Invariance Tests

A ML-CFA of fear appeal frequency, challenge evaluation, threat evaluation, and behavioral engagement, at T_1 and T_2, showed an acceptable fit to the data, $\chi^2(424) = 826.22$, p <.001, RMSEA = .026, $SRMR_W$ = .036, $SRMR_B$ = .086, CFI = .957, TLI = .944. In order to estimate latent bivariate correlations (see Table 2), mathematics grades and sociodemographic covariates (gender, age, and FSM), were added to the model. A ML-CFA showed an acceptable fit to the data, $\chi^2(539) = 1052.03$, p <.001, RMSEA = .026, $SRMR_W$ = .032, $SRMR_B$ = .085, CFI = .951, TLI = .933. More detail regarding the specification of the ML-CFAs can be found in the Supplementary Materials.

Measurement invariance tests for fear appeals and their evaluation, and behavioral engagement, were conducted as a precondition for modeling structural relations over time (Widaman & Reise, 1997). Fear appeals and their evaluation showed strict measurement invariance. Behavioral engagement demonstrated partial scalar invariance when the intercept for one between-level item was freed. Partial scalar invariance is sufficient for the longitudinal modeling of constructs (Widaman et al., 2010). Additional detail for the testing of measurement invariance is presented in the Supplementary Materials.

Multi-level Structural Equation Modeling

A ML-SEM was used to test the hypothesized paths shown in Figure 1. This model showed an acceptable fit to the data, $\chi^2(558) = 1138.34$, p <.001, RMSEA = .026, $SRMR_W$ = .035, $SRMR_B$ = .098, CFI = .943, TLI = .925, and so we proceeded to inspect path coefficients (see Table 3).

Within-Level Portion of the Model.

T_1 fear appeal frequency positively predicted T_2 challenge ($\beta = .11$, $p = .002$), but not T_2 threat ($\beta = -.02$, $p = .53$) evaluation, over and above the variance accounted for by autoregressive relations (T_1 challenge: $\beta = .52$, p <.001; T_1 threat: $\beta = .56$, p <.001) and Year 9 mathematics grade (T_2 challenge: $\beta = .03$, $p = .40$; T_2 threat: $\beta = -.10$, p =.004). T_2

behavioral engagement was positively predicted by T_1 challenge (β = .14, *p* = .003), but not T_1 threat (β = .02, *p* = .60), over and above the variance accounted for by the autoregressive relation (T_1 behavioral engagement: β = .39, *p* <.001). T_1 fear appeal frequency (β = .02, *p* = .67) and Year 9 mathematics grade (β = .01, *p* =.74) did not predict T_2 behavioral engagement. GCSE mathematics grade was predicted positively by T_2 behavioral engagement (β = .20, *p* <.001), and negatively by T_2 threat (β = -.21, *p* <.001), but not T_2 fear appeals (β = -.05, *p* =.24) or T_2 challenge (β = .03, *p* =.56), over and above the variance accounted for by Year 9 mathematics grade (β = .42, *p* <.001). Statistically significant coefficients are shown in Figure 2.

Between-Level Portion of the Model.

T_1 fear appeal frequency did not predict T_2 challenge (β = .13, *p* = .45) or T_2 threat (β = .05, *p* = .63) evaluation, over and above the variance accounted for by autoregressive relations (T_1 challenge: β = .60, *p* <.001; T_1 threat: β = .71, *p* <.001) and Year 9 mathematics grade (T_2 challenge: β = .18, *p* = .34; T_2 threat: β = -.09, *p* =.51). T_2 behavioral engagement was not predicted by T_1 challenge (β = .18, *p* = .55) or T_1 threat (β = -.01, *p* = .99), over and above the variance accounted for by the autoregressive relation (T_1 behavioral engagement: β = .04, *p* =.89). T_1 fear appeal frequency (β = .05, *p* =.18) and Year 9 mathematics grade (β = .20, *p* =.30) did not predict T_2 behavioral engagement. GCSE mathematics grade was negatively predicted by T_2 threat (β = -.34, *p* =.006), but not T_2 challenge (β = .04, *p* =.66), T_2 behavioral engagement (β = -.10, *p* =.25), or T_2 fear appeals (β = .05, *p* =.50), over and above the variance accounted for by Year 9 mathematics grade (β = .76, *p* <.001). Statistically significant coefficients are shown in Figure 3.

Indirect Paths

Indirect paths were tested by estimating 95% confidence intervals (CIs) around the unstandardized regression coefficients (see Table 4). CIs that do not cross zero are statistically significant (at p <.05). For the within-level portion of the model, indirect relations were shown from T_1 fear appeal frequency to T_2 behavioral engagement mediated by T_1/T_2 challenge evaluation (β = .020)[5] and from T_1 challenge evaluation to GCSE grade mediated by T_2 behavioral engagement (β = .035). In addition to these a priori theorized indirect paths, indirect relations between Year 9 mathematics grades and GCSE mathematics grades were examined in a supplementary analysis. Relations were mediated by T_2 threat evaluation (β = .022), T_1 and T_2 threat evaluation (β = .020), T_1 and T_2 behavioral engagement (β = .014), and T_1 challenge evaluation and T_2 behavioral engagement (β = .003). At the between-level portion of the model, an indirect relation between Year 9 and GCSE achievement was mediated by threat evaluation (β = .152). None of the other indirect paths, in the within- or between-level portions of the model, were statistically significant (ps >.05).

Supplementary Analyses

Three supplementary analyses were performed. First, as noted above, following the sequence of direct paths, we examined indirect paths from Year 9 to GCSE mathematics grade. Second, to examine the possibility of contextual effects, as described earlier, we estimated new parameters to reflect the L2 effect after accounting for the L1 effect, that were standardized for total variance and an effect size (ES) comparable to Cohen's d (see Morin et al., 2014). The path from T_2 threat to GCSE mathematics grade showed a statistically significant moderate contextual effect (β = -.15, SE = .07, p = .03, ES = -.65) and the path from Year 9 mathematics grade to GCSE mathematics grade showed a large contextual effect approaching statistical significance (β = .26, SE = .13, p = .05, ES = 1.13). All other contextual paths were not statistically significant (ps >.05). Third, for exploratory purposes

[5] Following the two-wave design, the indirect relation from fear appeals to behavioral engagement was estimated from the direct path from T_1 fear appeal frequency to T_2 challenge evaluation (path $a1$) and from T_1 challenge evaluation to T_2 behavioural engagement (path $b1$).

we examined the possibility that T_1 L2 fear appeals may have moderated the L1 relations between T_1 fear appeals and T_2 challenge/threat evaluation, and between T_1 challenge/threat evaluation and T_2 behavioral engagement. That is, L1 relations may have been amplified in classes where the teacher made more frequent fear appeals. All cross-level interactions were, however, not statistically significant (ps >.05).

Discussion

The aim of the study was to examine indirect relations from teacher fear appeals, i.e., messages about the importance of avoiding failure in the context of high-stakes school exit qualifications in mathematics, to achievement, mediated by fear appeal evaluations and behavioral engagement. The within-level portion of the model showed T_1 fear appeals predicted greater T_2 challenge evaluation, T_1 challenge evaluation predicted greater T_2 behavioral engagement, and T_2 behavioral engagement predicted a higher GCSE mathematics grade. Small positive indirect relations were shown between fear appeals and behavioral engagement, mediated by challenge evaluation, and between challenge evaluation and GCSE mathematics grade mediated by behavioral engagement. In addition, indirect relations between Year 9 mathematics and GCSE mathematics grades were mediated by T_1 and T_2 threat evaluation, T_1 and T_2 behavioral engagement, and T_1 challenge evaluation and T_2 behavioral engagement. The between-level portion of the model showed indirect relations between Year 9 and GCSE mathematics grades were mediated by T_1 and T_2 threat evaluation.

Fear Appeals and their Evaluation

Fear appeals were theorized as prompts for students to evaluate the perceived importance of their forthcoming high-stakes qualification and chances of success or failure. A greater frequency would, therefore, result in a stronger challenge evaluation if a student valued their examinations and anticipated success, or a threat evaluation if a student valued their examinations and anticipated failure (Putwain et al., 2021). A novel element of the present study was the inclusion of class mean-centered fear appeals in the within-level portion of the model. We conceptualized class mean-centered fear appeals as indicating the extent to which a student was attending to teacher messages; a higher score would indicate a student was more attentive to fear appeals than classmates.

The within-level portion of the model showed positive concurrent (within-wave) relations between fear appeals and challenge/threat evaluation. Notwithstanding the lack of temporal separation between fear appeals and evaluation here (precluding a directional conclusion), these positive relations are consistent with the proposition that fear appeals act as a prompt for challenge/threat evaluation. Between waves, T_1 fear appeals positively predicted T_2 challenge; T_1 fear appeals were unrelated to T_2 threat evaluation.

Previous studies have shown positive relations between fear appeals and challenge/threat evaluation when measured at the same wave (Putwain et al., 2016), and between waves after a four-month interval (Putwain et al., 2017). The latter study did not, however, control for concurrent or auto-lagged relations across waves. Findings of the present study were not entirely in keeping with these previous findings as fear appeals were only related to subsequent challenge evaluation. The present study addressed methodological limitations of the aforementioned studies (i.e., temporal separation of fear appeals and evaluations while controlling for concurrent and auto-lagged relations in a multi-level model).

There may be different reasons why a student may be more or less attentive to fear appeals than classmates. Students with high test anxiety could, for instance, attend more closely due to fear appeals as a signal for threat (Putwain & Best, 2012). Equally, students could attend more closely out of high motivation (see Acee et al., 2018). Given that a challenge evaluation has been shown in previous studies to be associated with adaptive achievement-related belief, emotion, and behavior (see Putwain et al., 2021), it is plausible in the present study that students were attending more closely out of high motivation or a similarly adaptive construct like academic buoyancy or adaptability (see Martin, 2013). For example, a student with a strong aspiration to continue academic study (utility or extrinsic value in EVT/CVT respectively; see Pekrun, 2018; Wigfield et al., 2016;) may attend more closely to fear appeals due to their personal significance (i.e., the consequential implications of success and failure).

If students were attending to fear appeals more closely due to adaptive motivation this could explain why fear appeals were not related to threat evaluation. It cannot be discounted, however, that fear appeals might prompt more proximal threat evaluation that, unlike challenge evaluation, does not persist over time. In addition, the shared concurrent variance between fear appeals and challenge/threat evaluation could result in insufficient unique variance to predict subsequent threat evaluation; all things being equal a higher challenge evaluation implies lower threat evaluation and vice versa (also see Nicholson et al., 2019). As it stands, however, results provide partial support for *Hypothesis 1* for the within-level portion of the model. While attending to a fear appeal may be a necessary condition for a threat evaluation (after all, it is a threat evaluation *of a fear appeal*), it is also possible that threat evaluation is more strongly influenced by other student-centered factors such as prior achievement (as shown in the present study), value of achievement and competence beliefs (e.g., Putwain & Symes, 2016; Symes & Putwain, 2016).

The difference between the within- and between-level portions of the model was striking. In the between-level portion of the model, there were no concurrent or time-lagged relations between class-aggregated fear appeals and challenge/threat evaluation. This finding runs contrary to the theoretical proposition that fear appeals prompt evaluations and previous studies showing positive relations between concurrent challenge/threat evaluations and fear appeal frequency (Putwain, Nicholson, et al., 2016, Putwain et al., 2017). These unexpected non-statistically significant findings cannot be attributed to teachers using fear appeals infrequently. In keeping with previous studies using teacher self-reports or aggregated student observer-reports (Putwain, Nicholson, et al., 2016, Putwain et al., 2017), teachers were reported to use fear appeals close to the mid-point of the response scale ("sometimes"; see Table 1). Similarly, findings cannot be attributed to low L2 reliability of aggregated student reports for teacher fear appeals; ICC_2 statistics were .80 and .86 for T_1 and T_2, respectively. *Hypothesis 1* was not supported for the between-level portion of the model.

It is possible that fear appeals at the class-level were not prompting evaluations. That is, irrespective of the frequency of fear appeals, class-level evaluations depend on other factors such as achievement (in the present study higher achievement was related to lower threat and higher challenge evaluation) and classroom climate (e.g., instructional, relational, and competence support). One would expect higher instructional and relational support to result in higher challenge and lower threat evaluation (e.g., Pino-James et al., 2019; Shernoff et al., 2017). Another possibility is a high degree of student heterogeneity in message evaluations within a single class. That is, students responded differently to teacher messages, such as fear appeals. If some students responded with greater challenge and others with lower challenge evaluaton, the net result, when aggregated, is that student differences partial each other out. Thus, we do not find overall levels of challenge and threat evaluations higher in classrooms where the teacher makes frequent fear appeals, not because the fear appeals do not function as a prompt. Rather, in keeping with findings from the within-level portion of the model, it is because the teacher message, although made to a whole class, is evaluated and responded to differently by students within that class.

Finally, it was notable that there were negative class-level relations between Year 9 mathematics achievement and subsequent use of fear appeals by teachers. It has been shown previously that teachers use fear appeals more frequently when they judge classes to be low in engagement (Putwain, Nakhla et al., 2017). The present study is the first to show that teachers also use more frequent fear appeals in lower achieving classes and supports the conceptualization of fear appeals as a teacher strategy to raise engagement and ultimately achievement.

A general point about the between-level portion of the model that applies to relations between fear appeal frequency and evaluation, as well as the sections that follow, concerning evaluation and achievement, is whether the non-significant relations might have resulted from an insufficiently powered model. General guidance suggests a minimum of 50 clusters (i.e., classes for the present study) under optimal conditions and ideally 100 (Hox & Maas, 2001). While this might seem to rule out the possibility of an underpowered model (there were 105 classes in the present study), ICC sizes, unequal cluster size and sampling, path sample sizes, and model complexity, can all influence the detection of significant paths in ML-SEM (McNeish, 2017).

We did not estimate paths from T_1 challenge/threat evaluation to T_2 fear appeal frequency. This was partly as bidirectional relations between fear appeals and evaluations were not a specific research question for the present paper and partly to avoid over-complicating an already complex model. It is notable, however, that bivariate correlations between T_1 challenge and threat evaluation and T_2 fear appeal frequency in the within-portion of the model were negligible (rs = .03 and .08, respectively), suggesting that evaluations were unrelated to subsequent attention to fear appeals. In the between-level portion of the model, T_1 challenge evaluation was negatively related to T_2 fear appeal frequency (r = -.15), although this was not statistically significant. T_1 threat evaluation was positively and significantly related to T_2 fear appeal frequency (r = .25); when classes are more likely to evaluate fear appeals as a threat, teachers use them more frequently. Although somewhat counter-intuitive, given the negative outcome associated with a threat evaluation, it does tally with previous findings (Putwain & von der Embse, 2018). Some teachers may believe classes can be motivated by fear of failure and hence use fear appeals more frequently in those classes that respond in this way, namely a threat response, to fear appeals.

Fear Appeals, Evaluations, and Engagement

It was hypothesized that a challenge evaluation would be positively, and threat evaluation negatively, related to behavioral engagement. Consistent with this proposition, the within-level portion of the model showed positive concurrent relations between challenge evaluation and behavioral engagement and negative concurrent relations between threat evaluation and behavioral engagement. Over time, T_1 challenge evaluation was positively related to T_2 behavioral engagement. T_2 threat evaluation, however, was unrelated to T_2 behavioral engagement. Thus, *Hypothesis 2* was partially supported at the within-level; students who make a challenge evaluation (i.e., they value GCSE mathematics grade and believe that with effort they can achieve their target grade) engage in greater on-task behavior in lessons six to seven months later.

The question is why only challenge and not threat evaluation was related to subsequent engagement. Previous studies have shown challenge evaluation to relate positively, and threat evaluation negatively, to concurrent behavioral engagement (Putwain et al., 2017; Putwain, Nakhla, et al., 2018) within waves. By contrast, Nicholson and Putwain, (2020) showed threat evaluation was negatively related to *emotional* rather than *behavioral* engagement, with and across waves (Nicholson & Putwain, 2020). Threat evaluation may, therefore, be more strongly related with subsequent achievement-related emotions than behavior. The weaker and/or non-significant relations with behavior could be an artifact of the dual effect of threat to increase anxiety but also motivate action to reduce anxiety (Eysenck et al., 2007; Pekrun, 2006). Motivated action can be protective when it reduces the threat, such as increased effort with study to avoid failure, but defensive when used to control anxiety, such as cognitive and behavioral avoidance resulting in less effort (e.g., Covington, 2009; Maloney & Lapinski, 2011).

In the between-level portion of the model, positive concurrent relations were shown between challenge evaluation and behavioral engagement; over time, however, challenge evaluation was unrelated to subsequent behavioral engagement. Threat evaluation was unrelated to concurrent or subsequent behavioral engagement. Findings are partly consistent with Putwain et al. (2017) who showed positive concurrent relations between challenge evaluation and behavioral engagement, and negative between threat evaluation and behavioral engagement. Challenge evaluation was most likely unrelated to behavioral engagement over time partly as a result of strong concurrent shared variance withub evaluation and behavioral engagement leaving insufficient unique variance. *Hypothesis 2* was not supported at the between-level.

Fear Appeals, Evaluations, Engagement, and Mathematics Achievement

We hypothesized that behavioral engagement would be related to subsequent GCSE grade and relations from fear appeals would be indirect and mediated through their evaluations and behavioral engagement. As expected, T_2 behavioral engagement positively predicted GCSE mathematics grade in the within-level portion of the model, beyond the variance accounted for by Year 9 mathematics examination grade. This finding replicates those shown in numerous previous studies of elementary (e.g., Dotterer & Lowe, 2011; Hughes & Kwok, 2007) and secondary (e.g., Martin & Liem, 2010; Wang & Holcombe, 2010) school students. In addition, as theorized, positive indirect relations were shown for the within-level portion of the model between fear appeals and behavioral engagement, mediated by challenge evaluation, and between challenge evaluation and GCSE grade, mediated by behavioral engagement. Although these paths were small, in relative terms, this is not unusual given that concurrent and autoregressive relations were controlled for (see Collie et al, 2015).

The finding that relations between fear appeals and achievement are mediated by challenge evaluation supports theorizing that fear appeals impact achievement indirectly (Putwain et al. 2021) and builds on previous studies showing indirect relations (Putwain & Symes, 2011; Putwain et al., 2017). Thus, students who attended to fear appeals and evaluated them as a challenge showed higher engagement, and students who showed higher engagement showed higher GCSE mathematics grades.

T_2 threat evaluation was also negatively related to subsequent GCSE grade in the within- and between-level portions of the model, and showed a contextual effect in the same direction. This finding is surprising as we had expected relations between threat evaluation and GCSE grade to be indirect and mediated by lower behavioral engagement. As we speculated above, the reason could be the absence of measuring emotional engagement. If emotional engagement (or alternately achievement emotions) had been included, the relation from threat evaluation to achievement would likely be indirect and mediated through lower emotional engagement or achievement emotions. While the negative relation between threat evaluation and GCSE achievement does not directly support our theorizing, it is consistent with the conceptualization that threat evaluation is a non-adaptive evaluation of fear appeals. The contextual effect of the path from T_2 threat evaluation to GCSE grade showed, after removing individual level variance, lower GCSE mathematics grades in classes where there is a higher average threat evaluation.

The indirect relations shown from Year 9 mathematics grade to GCSE mathematics grade showed lower threat evaluation was beneficial to subsequent achievement in a manner reminiscent of studies showing reciprocal relations between learning/classroom anxiety and achievement (e.g., Pekrun et al., 2017; Putwain et al., 2020). Thus, students with higher prior achievement, and who evaluated fear appeals as low in threat, received an incrementally higher GCSE grade than would have been expected from their prior achievement alone (and vice versa). This is a likely function of the high value and strong competence beliefs that comprise a low threat evaluation (Putwain & Symes, 2016; Symes & Putwain, 2016). The findings of the present study are unique in demonstrating negative relations between a threat evaluation and achievement at the classroom, as well as the student level. It is notable that stronger relations were observed at the classroom level which may be partly a result of students being placed in ability grouped classes for their GCSE programme of study at the beginning of Year 10. The contextual effect of the relations between Year 9 and GCSE mathematics grades showed that, after removing individual level variance, students in classes with higher average Year 9 mathematics grades achieved higher GCSEs grades in part due to lower threat evaluation.

Although the within-level indirect paths reported in Table 4 were small, this is to be expected in naturalistic studies with multiple waves of data collection that control for autoregressive paths (Collie et al., 2015). We would further add that not only were autoregressive paths controlled for, concurrent relations were accounted for also and there was a substantial interval between the two waves of data collection. This makes for an extremely thorough and robust test of indirect relations which will inevitably result in smaller coefficients. Larger indirect effects are often only found in complex naturalistic data when autoregressive effects have not been accounted for (Martin, 2011). The question is whether these indirect relations are substantively meaningful. Our position is that it is remarkable, given the rigorous design, that even small indirect relations were found.

Student attention to fear appeals may only result in small changes in subsequent behavioral engagement, and behavioral engagement to subsequent achievement, if considered in isolation. However, even small gains and losses can be meaningful and, if occurring at a grade boundary, could mean the difference between a pass and fail. Furthermore, we collected data over two waves separated by six to seven months. It is likely that iterative relations between fear appeals and their evaluations, engagement, and achievement, occur over shorter periods of time (e.g., days and weeks). Small relations could build in a cyclical fashion over successive iterations into more powerful effects.

T_2 behavioral engagement was unrelated to GCSE achievement for the between-level portion of the model, meaning that, overall, *Hypothesis 3* was partially supported. This finding was unexpected and appears counter-intuitive. How could those classes who engage, on average, more highly than others not achieve more highly? The answer may be partly a statistical artifact of the relatively large correlation between T_2 challenge evaluation and T_2 behavioral engagement reducing the unique variance of T_2 behavioral engagement to simultaneously predict GCSE mathematics grade. Such collinearity may have also contributed to the high standard error of the path from T_2 behavioral engagement to GCSE mathematics grade, relative to the standardized path coefficient. However, it is also possible that, since classes were grouped by ability, students in some high ability classes did not engage as highly as students in lower ability classes. Those students in lower ability classes may have felt the need for greater engagement to achieve target grades in contrast to those students who, rightly or wrongly, believed their ability mitigated the need for as much effort (e.g., Linnenbrink-Garcia et al., 2018; Ramos et al., 2021).

Limitations and Directions for Future Research

Despite the robust design and analysis used in the present study, and the novel contribution to the extant literature, there are a number of limitations to highlight. First, we focused on a single form of engagement, namely behavioral engagement. Other forms of engagement can be measured, such as cognitive, emotional, and psychological engagement (e.g., Appleton et al., 2006; Fredricks et al., 2011). Including alternative forms of engagement, as well as disaffection (see Skinner et al., 2009), could potentially reveal more nuanced relations than in the present study, for example, relations from threat evaluation to achievement mediated by emotional engagement or disaffection.

Second, our study contained only two waves of self-reported data collection. Although two waves are sufficient to test indirect relations (see Cole & Maxwell, 2003; Little et al., 2007), it meant that indirect relations between fear appeals and behavioral engagement could only be tested once. We had intended to collect a third wave of data in March 2020 (thereby enabling indirect relations between fear appeals and behavioral engagement to be tested twice), however the start of the COVID-19 pandemic in early 2020, resulting in English schools closing in March, meant this was not possible.

A third limitation, also resulting from the COVID-19 pandemic, was that we were unable to use GCSE examination grades in mathematics. It may be reasonably questioned whether fear appeals made by the teacher during a period of instruction when examinations were the expected mode of assessment (i.e., prior to schools closing in March 2020) would apply when examinations were not ultimately used. As the high stakes associated with GCSEs (i.e., awarded grades were still legitimately required to access post-secondary education, training, and employment opportunities) remained irrespective of whether the qualification was awarded on the basis of examinations or teacher estimates, we argue that teacher fear appeals were still of relevance. Furthermore, items used to measure the frequency of teacher fear appeals reference *failure,* rather than *examination failure* more specifically, hence could apply to different assessment formats. It is also germane to highlight that developments of the COVID-19 pandemic would not have interfered with teacher use of fear appeals, evaluations, and engagement over the two waves of data collection; these occurred prior to the first reported case. It is, however, unlikely that teacher estimated grades would be as accurate and consistent as an examination grade from an external awarding body.

Fourth, the present study did not consider success appeals and evaluations in tandem with fear appeals and evaluations. It may be that some teachers make both success and fear appeals whereas others may use fear appeals more frequently (or at least with some classes) than success appeals (and vice versa). The available evidence shows that students evaluate fear and success appeals in similar ways (Putwain & Symes, 2016; Symes & Putwain, 2016). It is possible that even if they were evaluated in similar ways, success appeals could lead to better outcomes than fear appeals, or there may be an interaction between message frame and personal characteristics. That is, some students may benefit more from success appeals and others from fear appeals.

Fifth, our study relied solely on student reports of teacher fear appeals and self-perceptions of fear appeal evaluations and behavioral engagement. Measurements of teacher behavior may be influenced by liking of, or rapport with, their mathematics teacher (Donker et al., 2021; Li et al., 2012). Measurements of self-perceptions may be influenced by self-protective or self-enhancement biases, and the limits of self-knowledge (Bollich et al., 2015). Future studies should consider the use of teacher self-reports of fear appeals to compliment those of student reports, and the use of informant reports, such as those of peers, parents, and teachers, for student evaluations of fear appeals and their behavioral engagement (Connelly & Ones, 2010). Of course, teacher self-reported behaviors and informant reports of others are also subject to potential biases (Leising et al., 2014). The triangulation and comparison of multiple sources of information could, however, enable a more nuanced understanding of teacher fear appeals and their possible consequences as well as how different data sources may influence findings.

Sixth, and finally, we used an inter-individual approach to modeling data in the present study based on individual differences between students' fear appeal evaluations, behavioral engagement, and so on. This approach, however, does not account for the realtime unfolding of processes within individual students and within classrooms (Schmitz, 2006; Schmitz & Skinner, 1993). Indeed, it could be argued that relations between fear appeal evaluations and outcomes are actually intra-individual in nature and inter-individual differences should only be considered once within-person processes have been established (Hamaker, 2012). An intra-individual approach could provide valuable insights into the stability and temporal dynamics of students' classroom experiences within lessons and from one lesson to the next (Vasalampi et al., 2021). Researchers may, therefore, wish to consider using experience sampling methods (see Malmberg, 2020) in future studies of fear appeals, their evaluation, and their outcomes.

Implications for Educational Practice

Students who evaluated fear appeals as a challenge showed greater behavioral engagement and those who evaluated fear appeals as a threat showed lower achievement. These findings imply that fear appeals are neither positive nor negative in themselves; it depends on how they are evaluated. Furthermore, as no between-level paths were shown in our ML-SEM between fear appeals and their evaluations, and subsequent engagement or achievement, our recommendation is not as straightforward as teachers using more or less frequent fear appeals. Indeed, it may be difficult for students not to be exposed to consequential messages about success or failure when examinations or qualifications are high-stakes such as the GCSE context of the present study.

The critical point is the type of evaluations that students are making. Teachers could be advised that, irrespective of how frequent fear appeals are made, there will be individual differences between students in how they are evaluated and responded to; some students will likely benefit, others will be likely hindered. Although we did not empirically address the differential antecedents of challenge and threat evaluation in the present study, it is known from previous studies that challenge evaluation follows expectations of success when examination outcomes are valued and threat evaluation follows expectations of failure (or uncertain expectations of success) when examination outcomes are valued (e.g., Putwain & Remedios, 2014; Putwain, Remedios, et al., 2015). In addition, the present study showed that challenge evaluation was more likely when prior achievement was higher and threat evaluation was more likely when prior achievement was lower. Indeed, prior achievement may have informed subsequent expectations of success and failure.

An effective approach may be to identify those students likely to respond with a threat evaluation. Bearing in mind the aforementioned possibility of intra-individual differences, it would also be useful to identify if there are particular times when students evaluate fear appeals with greater threat (such as immediately prior to examinations) than others. Educational interventions designed to build confidence in success can be achieved through strategy-focused feedback, subject mastery, and effort reinforcement (e.g., Linnenbrink-Garcia et al., 2016; Ramirez et al., 2018). Psychological interventions can be implemented to help students respond to threat with effort to reduce the likelihood of failure rather than avoidance to control anxiety (e.g., Putwain & Pescod, 2018).

It is also questionable the extent to which teachers may be able to effectively judge the private motivational beliefs of students (e.g., Urhahne et al., 2011). Flitcroft et al. (2017) showed that teachers were open to learning new ways to gain feedback on students' private self-perceptions in order to communicate the most effective message prior to high-stakes examinations. Methods to capture student voice with respect to their confidence and fears may, therefore, prove useful for teachers in this respect.

Conclusions

Using a ML-SEM, controlling for autoregressive and concurrent relations, we showed that students who attend more closely to messages from teachers that highlight the negative consequences of failure (fear appeals) were more engaged when fear appeals were evaluated as a challenge. In addition, achievement was lower in students who evaluated fear appeals as a threat. These findings show that fear appeals are per se, neither effective nor ineffective to achieving positive outcomes such as behavioral engagement and achievement. Rather, it is how the messages are evaluated that determines a more or less favorable outcome for the student. Accordingly, teachers are advised to identify those students likely to respond to fear appeals with a threat and consider the use of educational or psychological intervention to reduce this likelihood (i.e., building confidence and reducing anxiety control through avoidance). As teachers may experience difficulty in judging those students likely to respond with a threat evaluation, methods to access student voice may be beneficial.

References

Acee, T.W., Weinstein, C.E., Hoang T.V., & Flaggs, D.A. (2018). Value Reappraisal as a Conceptual Model for Task-Value Interventions. *The Journal of Experimental Education, 86*(1), 69-85.doi: 10.1080/00220973.2017.1381830

Anders, J., Dilnot, C., Macmillan, L., & Wyness, G. (2020). *Grade expectations: How well can we predict future grades based on past performance? (Working Paper No. 20-14)*. Centre for Education Policy and Equalising Opportunities, UCL Institute of Education.

Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of School Psychology*, *44*(5), 427–445. https://doi.org/10.1016/j.jsp.2006.04.002.

Banks, J., & Smyth, E. (2015). 'Your whole life depends on it': academic stress and high-stakes testing in Ireland. *Journal of Youth Studies*, *18*(5), 598-616. https://doi.org/10.1080/13676261.2014.992317

Bishop, J. H., & Mane, F. (2001). The impacts of minimum competency exam graduation requirements on high school graduation, college attendance and early labor market success. *Labour Economics*, *8*(2), 203-222. https://doi.org/10.1016/S0927-5371(01)00025-2

Bollich, K. L., Rogers, K. H., & Vazire, S. (2015). Knowing more than we can tell: People are aware of their biased self-perceptions. *Personality and Social Psychology Bulletin, 41,* 918–929. http://doi.org/10.1177/0146167215583993

Breakwell, G.M., Fino, E., & Jaspal, R. (2021). The COVID- 19 preventive behaviors index: Development and validation in two samples from the United Kingdom. *Evaluation and the Health Professions, 44*(1), 77-86. https://doi.org/10.1177/0163278720983416

Carnoy, M. (2005). Have State Accountability and High-Stakes Tests Influenced Student Progression Rates in High School? *Educational Measurement: Issues and Practice, 24*(4), 19-31. https://doi.org/10.1111/j.1745-3992.2005.00020.x

Coertjens, L., Donche, V., De Maeyer, S., Vanthournout, G., & Van Petegem, P. (2017). To what degree does the missing-data technique influence the estimated growth in learning strategies over time? A tutorial example of sensitivity analysis for longitudinal data. *PLoS ONE, 12*(9), e0182615. https://doi.org/10.1371/journal.pone.0182615

Cole, D. A., Ciesla, J. A., & Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychological Methods, 12*(3), 381–398. https://doi.org/10.1037/1082-989X.12.4.381

Cole, D. A., & Maxwell, S. E. (2003). Testing Mediational Models With Longitudinal Data: Questions and Tips in the Use of Structural Equation Modeling. *Journal of Abnormal Psychology, 112*(4), 558–577. https://doi.org/10.1037/0021-843X.112.4.558

Collie, R.J., Martin, A.J., Malmberg, L-E., Hall, J., & Ginns, P. (2015). Academic buoyancy, student’s achievement, and the linking role of control: A cross-lagged analysis of high school students. *British Journal of Educational Psychology, 85*(1), 113–130. https://doi.org/10.1111/bjep.12066

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin*, *136*(6),1092–1122. http://doi.org/10.1037/a0021212

Covington, M. (2009). Self-worth theory: Retrospects and prospects. In K. R. Wentzel, & A. Wigfield (Eds.). *Handbook of motivation at school* (pp. 141–170). Routledge.

Donker, M. H., Vemde, L. van, Hessen, D. J., Gog, T. van, & Mainhard, T. (2021). Observational, student, and teacher perspectives on interpersonal teacher behavior: Shared and unique associations with teacher and student emotions. *Learning and Instruction*, *73,* 101414. https://doi.org/10.1016/j.learninstruc.2020.101414

Dotterer, A. M., & Lowe, K. (2011). Classroom context, school engagement, and academic achievement in early adolescence. *Journal of Youth and Adolescence, 40*, 1649–1660. https://doi.org/10.1007/s10964-011-9647-5

Department for Education (2019). *Schools, pupils, and their characteristics: January 2019*. Her Majesty's Stationary Office.

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: attentional control theory. *Emotion, 7*(2), 336-353. https://doi.org/ 10.1037/1528-3542.7.2.336

Flitcroft, D., Woods, K., & Putwain, D. W. (2017). Developing school practice in preparing students for high-stake examinations in English and Mathematics. *Educational and Child Psychology*, *34*(3), 7–19. https://shop.bps.org.uk/educational-child-psychology-vol-34-no-3-september-2017-emergent-topics-in-educational-psychology

Folkman, S., & Lazarus, R. S. (1985). If it changes it must be a process: Study of emotion and coping during three stages of a college examination. *Journal of Personality and Social Psychology, 48*(1), 150–170. https://doi.org/10.1037/0022-3514.48.1.150

Fredricks, J., McColskey,W., Meli, J., Mordica, J., Montrosse, B., & Mooney, K. (2011). *Measuring student engagement in upper elementary through high school: a description of 21 instruments*. (Issues & Answers Report, REL 2011–No. 098). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. Retrieved from http://ies.ed.gov/ncee/edlabs.

Green, J., Martin, A.J., & Marsh, H.W. (2007). Motivation and engagement in English, mathematics and science high school subjects: Towards an understanding of multidimensional domain specificity. *Learning and Individual Differences, 17*(3), 269-279. http://dx.doi.org/10.1016/j.lindif.2006.12.003

Hamaker, E. L. (2012). Why researchers should think within-Person: A pragmatic rationale. In M. R. Mehl & T. S. Conner (Eds.) *Handbook of Research Methods for Studying Daily Life* (pp. 43–61). Guilford.

Hamre, B., Hatfield, B., Pianta, R., & Jamil, F. (2014). Evidence for general and domain-specific elements of teacher-child interactions: Associations with preschool children's development. *Child Development*, *85*(3), 1257–74. https://doi.org/10.1111/cdev.12184

Hattie, J. (2009). *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement*. Routledge.

Heath, A. F., Rothon, C., & Kilpi, E. (2008). The second generation in Western Europe: Education, unemployment, and occupational attainment. *Annual Review of Sociology*, *34*, 211-235. https://doi.org/10.1146/annurev.soc.34.040507.134728

Hofmann, D. A. (2002). Issues in multilevel research: Theory development, measurement, and analysis. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology* (pp. 247–274). Blackwell.

Howe, C., Hennessy, S., Mercer, N., Vrikki, M., & Wheatley, L. (2019). Teacher-student dialogue during classroom teaching: Does it really impact on student outcomes? *Journal of the Learning Sciences*, *28*(4-5), 462–512. https://doi.org/10.1080/10508406.2019.1573730

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications (2nd ed.)*. Routledge

Hox, J. J., & Maas, C. J. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling, 8*(2), 157–174. https://doi.org/10.1207/S15328007SEM0802_1

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hughes, J., & Kwok, O. M. (2007). Influence of student-teacher and parent-teacher relationships on lower achieving readers' engagement and achievement in the primary grades. *Journal of Educational Psychology, 99*(1), 39–51. doi:10.1037/0022-0663.99.1

Kearney, P., Plax, T. G., Richmond, V. P., & McCroskey, J. C. (1985). Power in the classroom III: Teacher communication techniques and messages. *Communication Education. 34*(1), 19–28. https://doi.org/10.1080/03634528509378579

Kuppens, P., & Tong, E. M. (2010). An appraisal account of individual differences in emotional experience. *Social and Personality Psychology Compass, 4*(12), 1138-1150. https://doi.org/10.1111/j.1751-9004.2010.00324.x

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9*, 231–251. https://doi.org/10.1007/s10984-006-9015-7

Lazarus, R. S. (2001). Relational meaning and discrete emotions. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 37–67). Oxford University Press.

Lazarus, R. S. and Folkman, S. (1984). *Stress, appraisal, and coping*. Springer.

Leising, D., Gallrein, A. M. B., & Dufner, M. (2014). Judging the behavior of people we know: Objective assessment, confirmation of preexisting views, or both? *Personality & Social Psychology Bulletin*, *40*, 153–63. http://doi.org/10.1177/0146167213507287

Li, Y., Hughes, J. N., Kwok, O.-m., & Hsu, H.-Y. (2012). Evidence of convergent and discriminant validity of child, teacher, and peer reports of teacher–student support. *Psychological Assessment, 24*(1), 54–65. https://doi.org/10.1037/a0024481

Linnenbrink-Garcia, L., Patall, E. A., & Pekrun, R. (2016). Adaptive motivation and emotion in education: Research and principles for instructional design. *Policy Insights from the Behavioral and Brain Sciences, 3*(2)*,* 228-236. https://doi.org/ 10.1177/2372732216644450

Linnenbrink-Garcia, L., Wormington, S. V., Snyder, K. E., Riggsbee, J., Perez, T., Ben-Eliyahu, A., & Hill, N. E. (2018). Multiple pathways to success: An examination of integrative motivational profiles among upper elementary and college students. *Journal of Educational Psychology, 110*(7), 1026–1048. https://doi.org/10.1037/edu0000245

Little, T. D., Preacher, K. J., Card, N. A., & Selig, J. P. (2007). New developments in latent variable panel analyses of longitudinal data. *International Journal of Behavioral Development, 31*(4), 357-365. https://doi.org/10.1177/0165025407077757

Long, R., & Bolton, P. (2017). *English Baccalaureate (House of Commons Briefing Paper: 06045)*. Her Majesty's Stationary Office.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modelling. *Contemporary Educational Psychology, 34*(2), 120-131. https://doi.org/10.1016/j.cedpsych.2008.12.001

Maguire, S. (2010). 'I just want a job'–what do we really know about young people in jobs without training? *Journal of Youth Studies*, *13*(3), 317-333. https://doi.org/10.1007/s10984-006-9015-71

Malmberg, L.-E. (2020). Intraindividual Structural Equation Models for Learning Experiences. *International Journal of Research and Methods in Education 43*(4), 413-430. https://doi.org/10.1080/1743727X.2020.1793939

Maloney, E. K., Lapinski, M. K., Witte, K. J. S., and Compass, P. P. (2011). Fear appeals and persuasion: A review and update of the extended parallel process model. *Social and Personality Psychology Compass, 5*(4), 206-219. https://doi.org/10.1111/j.1751-9004.2011.00341.x

Marsh, H.W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., et al. (2012). Classroom climate and contextual effects: conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*(2), 106-124. http://dx.doi.org/10.1080/00461520.2012.670488

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B.,et al. (2009). Doubly-latent models of school contextual effects: integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research, 44*(6), 764-802. http://dx.doi.org/10.1080/00273170903333665

Martin, A. J. (2011). Prescriptive statements and educational practice: What can Structural Equation Modeling (SEM) offer? *Educational Psychology Review, 23*(2), 235–244. https://doi.org/10.1007/s10648-011-9160-0

Martin, A.J. (2013). The Personal Proficiency Network: Key self-system factors and processes to optimize academic development. In D.M. McInerney, H.W. Marsh., R.G. Craven, & F. Guay (Eds). *Theory driving research: New wave perspectives on self-processes and human development* (pp. 251–286). Information Age Publishing.

Martin, A. M., & Liem, G. A. D. (2010). Academic personal bests (PBs), engagement, and achievement: A cross-lagged panel analysis. *Learning and Individual Differences, 29*(3), 265–270. https://doi.org/10.1016/j.lindif.2010.01.001

McNeish, D. (2017). Multilevel mediation with small samples: A cautionary note on the multilevel structural equation modeling framework. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(4), 609-625. https://doi.org/10.1080/10705511.2017.1280797

Mercer, N., & Littleton, K. (2007). *Dialogue and the development of children's thinking: A sociocultural approach*. Routledge.

Monagas, E. S., & Putwain, D. W., Leon, J., Núñez, J. L. & . Loro, J. F. (2021). Do motivational messages predict motivation to learn and performance? A theory integration approach. *British Journal of Educational Psychology*. Manuscript submitted for publication.

Morin, A.J.S., Marsh, H.W., Nagengast, B., & Scalas, L.F. (2014). Doubly latent multilevel analyses of classroom climate: An illustration. *The Journal of Experimental Education, 82*(2), 143-167. http://dx.doi.org/10.1080/00220973.2013.769412

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8^{th} ed.). Muthén & Muthén.

Nicholson, L. J., Putwain, D. W., Nakhla, G., Porter, B., Liversidge, A., & Reece, M. (2019). A person-centered approach to students' evaluations of perceived fear appeals and their association with engagement. *The Journal of Experimental Education, 87*(1), 139-160. https://doi.org/10.1080/00220973.2018.1448745

Nicholson, L. J., & Putwain, D. W. (2020). A cross-lagged panel analysis of fear appeal appraisal and student engagement. *British Journal of Educational Psychology*, *90*(3), 830-847. https://doi.org/10.1111/bjep.12334

Office of Qualifications and Examinations Regulation. (2016). *Setting the grade standards of new GCSEs in England.* Her Majesty's Stationary Office.

Office of Qualifications and Examinations Regulation (2020). *Summer 2020 grades for GCSE, AS and A level, Extended Project Qualification and Advanced Extension Award in maths: Guidance for teachers, students, parents and carers*. Her Majesty's Stationary Office.

Padgett, R.H., & Morgan, G.B. (2021) Multilevel CFA with ordered categorical data: A simulation study comparing fit indices across robust estimation methods. *Structural Equation Modeling: A Multidisciplinary Journal, 28*(1), 51-68. https://doi.org/10.1080/10705511.2020.1759426

Paulus, D. L., & Vazire, S. (2010). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224– 239). Guilford Press.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18*(4), 315–341. https://doi.org/10.1007/s10648-006-9029-9

Pekrun, R. (2018). Control-value theory: A social-cognitive approach to achievement emotions. In G. A. D. Liem & D. M. McInerney (Eds.), *Big theories revisited 2: A* *volume of research on sociocultural influences on motivation and learning* (pp. 162-190). Information Age Publishing.

Pekrun, R., & Perry, R. P. (2014). Control-value theory of achievement emotions. In R. Pekrun & L. Linnenbrink-Garcia (Eds.), *International handbook of emotions in education* (Vol. 1–Book, Section, pp. 130–151). Routledge.

Pekrun, R., Lichtenfeld, S., Marsh, H. W., Murayama, K., & Goetz, T. (2017). Achievement emotions and academic performance: Longitudinal models of reciprocal effects. *Child Development, 88,* 1653-1670. https://doi.org/10.1111/cdev.12704

Perryman, J. (2006). Panoptic performativity and school inspection regimes: Disciplinary mechanisms and life under special measures. *Journal of Education Policy, 21*(2), 147-161. https://doi.org/10.1080/02680930500500138

Perryman, J., Ball, S., Maguire, M., & Braun, A. (2011). Life in the pressure cooker–School league tables and English and mathematics teachers' responses to accountability in a results driven era. *British Journal of Educational Studies*, *59*(2), 179-195. https://doi.org/10.1080/00071005.2011.578568

Piccolo, L. R., Giacomoni, C. H., Julio-Costa, A., Oliveira, S., Zbornik, J., Haase, V. G., & Salles, J. F. (2017). Reading anxiety in L1: Reviewing the concept. *Early Childhood Education Journal*, *45*(4), 537-543. https://doi.org/10.1007/s10643-016-0822-x

Pino-James, N., Shernoff, D. J., Bressler, D. M., Larson, S. C., & Sinha, S. (2019). Instructional interventions that support student engagement: An international perspective. In J. A. Fredricks, A. L. Reschly, & S. L. Christenson (Eds.), *Handbook of Student Engagement Interventions* (pp. 103-119). Elsevier.

Preckel, F., & Brüll, M. (2010). The benefit of being a big fish in a big pond: Contrast and assimilation effects on academic self-concept. *Learning and Individual Differences, 20*(5), 522-531. https://doi.org/10.1016/j.lindif.2009.12.007

Putwain, D.W. (2009). Assessment and examination stress in Key Stage 4. *British Educational Research Journal, 35*(3), 391-411. https://doi.org/10.1080/01411920802044404

Putwain, D. W., & Best, N. (2011). Fear appeals in the primary classroom: Effects on test anxiety and test grade. *Learning and Individual Differences*, *21*(5), 580-584. https://doi.org/10.1016/j.lindif.2011.07.007

Putwain, D.W., Nakhla, G., Liversidge, A., Nicholson, L.J., Porter, B., & Reece, M. (2017). Teachers use of fear appeals prior to a high-stakes examination: Is frequency linked to perceived student engagement and how do students respond? *Teaching and Teacher Education, 61*, 73–83. https://doi.org/10.1016/j.tate.2016.10.003

Putwain, D.W., Nicholson, L.J., Nakhla, G., Reece, M., Porter, B., & Liversidge, A. (2016). Fear appeals prior to a high-stakes examination can have a positive or negative impact on engagement depending on how the message is appraised. *Contemporary Educational Psychology 44-45*, 21–31. https://doi.org/10.1016/j.cedpsych.2015.12.001

Putwain, D.W., & Pescod, M. (2018). Is reducing uncertain control the key to successful test anxiety for Secondary school students? Findings from a randomized control trial. *School Psychology Quarterly, 33*(2), 283-292. https://doi.org/10.1037/spq0000228

Putwain, D.W. & Remedios, R. (2014) Fear appeals used prior to a high-stakes examination: What makes them threatening? *Learning and Individual Differences, 36*, 145–151. https://doi.org/10.1016/j.lindif.2014.10.018

Putwain, D.W., Remedios, R., & Symes, W. (2015). Fear appeals used prior high-stakes examinations: Why are they appraised as threatening and do they impact on subjective task value? *Learning and Instruction, 40*, 21–28. https://doi.org/10.1016/j.learninstruc.2015.07.007

Putwain, D.W., Remedios, R., & Symes, W. (2016). The appraisal of fear appeals as threatening or challenging: Frequency of use, academic self-efficacy and subjective value. *Educational Psychology, 36*(9), 1677–1697. https://doi.org/10.1080/01443410.2014.963028

Putwain, D. W., & Roberts, C. M. (2012). Fear and efficacy appeals in the classroom: The secondary teachers' perspective. *Educational Psychology*, *32*(3), 355-372. https://doi.org/10.1080/01443410.2012.659845

Putwain, D.W., & Symes, W. (2011). Classroom fear appeals and examination performance: facilitating or debilitating outcomes? *Learning and Individual Differences, 21*(2), 227-232. https://doi.org/10.1016/j.lindif.2010.11.022

Putwain, D.W. & Symes, W. (2014). Subjective value and academic self-efficacy: The appraisal of fear appeals used prior to a high-stakes test as threatening or challenging. *Social Psychology of Education, 17*(2), 229-248. https://doi.org/10.1007/s11218-014-9249-7

Putwain, D.W., & Symes, W. (2016). The appraisal of value-promoting messages made prior to a high-stakes mathematics examination: The interaction of message-focus and student characteristics. *Social Psychology of Education, 19*(2), 325–343. https://doi.org/10.1007/s11218-016-9337-y

Putwain, D.W., Symes, W., Coxon, E., & Gallard, D. (2020). Attention bias in test anxiety: The impact of a test-threat congruent situation, presentation time, and approach-avoidance temperament. *Educational Psychology, 40*(6), 713-734. https://doi.org/10.1080/01443410.2020.1740653

Putwain, D.W., Symes, W., & McCaldin, T. (2019). Teacher use of loss-focused, utility value messages, prior to high-stakes examinations, and their appraisal by students. *Journal of Psychoeducational Assessment, 37*(2), 169-180. https://doi.org/10.1177/0734282917724905

Putwain, D. W., Symes, W., Nicholson, L. J., & Becker, S. (2018). Achievement goals, behavioural engagement, and mathematics achievement: A mediational analysis. *Learning and Individual Differences, 68,* 12–19. https://doi.org/10.1016/j.lindif.2018.09.006

Putwain, D.W., Symes, W., Nicholson, L.J., & Remedios, R. (2021). Teacher motivational messages prior to examinations: What are they, how are they evaluated, and what are their educational outcomes? In A.J. Elliot (Ed.), *Advances in motivational science* (pp. 63 – 103). Academic Press.

Putwain, D.W., Symes, W., & Remedios, R. (2016). The impact of fear appeals on subjective-task value and academic self-efficacy: The role of appraisal. *Learning and Individual Differences 51,* 307–313. https://doi.org/10.1016.j.lindif.2016.08.042

Putwain, D.W., Symes, W., & Wilkinson, H.M. (2017). Fear appeals, engagement, and examination performance: The role of challenge and threat appraisals. *British Journal of Educational Psychology, 87*(1), 16-31. https://doi.org/10.1011./bjep.12132

Putwain, D.W., & von der Embse, N. P. (2018). Teachers Use of Consequence and Timing Reminders: Links to Environmental and Personal Factors. *Social Psychology of Education, 21*(5), 1001-2019. https://doi.org/10.1007/s11218-018-9448-8

Putwain, D.W., Wood, P., & Pekrun, R. (2020). Achievement emotions and academic achievement: Reciprocal relations and the moderating influence of academic buoyancy. *Journal of Educational Psychology*. Advance online publication. https://doi.org/10.1037/edu0000637

Putwain, D.W., & Woods, K. (2016). The scare tactic: Does it work? Motivating students for tests and exams. *DECP Debate, 160*, 31–38.

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework forassessing multilevel mediation. *Psychological Methods*, *15*(2), 209-233. https://doi.org/10.1037/a0020141

Ramos, Al., Lavrijsen, J., Soenens, B., Vansteenkiste, M., Sypré, S., & Verschueren, K. (2021). Profiles of maladaptive school motivation among high-ability adolescents: A person-centered exploration of the motivational Pathways to Underachievement model. *Journal of Adolescence, 88*, 146-161. https://doi.org/10.1016/j.adolescence.2021.03.001

Ramirez, G., Shaw, S. T., & Maloney, E. A. (2018). Math anxiety: Past research, promising interventions, and a new interpretation framework. *Educational Psychologist, 53*(3), 145-164. https://doi.org/10.1080/00461520.2018.1447384

Reyes, M. R., Brackett, M. A., Rivers, S. E., White, M., & Salovey, P. (2012). Classroom Emotional Climate, Student Engagement, and Academic Achievement. *Journal of Educational Psychology*, 104(3), 700-712. https://doi.org/10.1037/a0027268

Richmond, V. P., & McCroskey, J. C. (1992). Increasing teacher influence through immediacy. In V. P. Richmond and J. C. McCroskey (Eds.), *Power in the classroom: Communication, control and concern* (pp. 101–119). Erlbaum.

Roberts, K. (2004). School-to-work transitions: Why the United Kingdom's educational ladders always fail to connect. *International Studies in Sociology of Education, 14*(3), 203-216. https://doi.org/10.1080/09620210400200126

Roberts, N. & Abreu, L. (2016). *School inspections in England: Ofsted (Briefing Paper 07091)*. London: House of Commons Library. Retrieved from: http://researchbriefings.files.parliament.uk/documents/SN07091/SN07091.pdf

Rothman, A. J., & Salovey, P. (1997). Shaping perceptions to motivate healthy behavior: the role of message framing. *Psychological bulletin*, *121*(1), 3-19. https://doi.org/10.1037/0033-2909.121.1.3

Ruiter, R.A.C., Kessels, L.T.E., Peters, G-J., Y., & Kok, G. (2014). Sixty years of fear appeals research: Current state of the evidence. *International Journal of Psychology, 49*(2), 63–70. https://doi.org/10.1002/ijop.12042

Santaga, E., & Putwain, D.W., Leon, J. Núñez, J.L. & . Loro, J.F. (2021). Do motivational messages predict motivation to learn and performance? *Journal of Psychodidactics*. Advance online publicaton. https://doi.org/10.1016/j.psicod.2021.11.001

Schmitz, B. (2006). Advantages of studying processes in educational research. *Learning and Instruction, 16*, 433–449. https://doi.org/10.1016/j.learninstruc.2006.09.004

Schmitz, B., & Skinner, E. (1993). Perceived control, effort, and academic performance: Interindividual, intraindividual, and multivariate time-series analyses. *Journal of Personality and Social Psychology, 64*(6), 1010–1028. https://doi.org/10.1037/0022-3514.64.6.1010

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, R.J. (2014). The history of qualifications and the role of competition. In H.G. Sahlgren (Ed.), *Tests worth teaching to: Incentivising quality in qualifications and accountability* (pp. 6-27). The Centre for Market Reform of Education

Shernoff, D. J., Ruzek, E. A., & Sinha, S. (2017). The influence of the high school classroom environment on learning as mediated by student engagement. *School Psychology International*, *38*(2), 201-218. https://doi.org/10.1177/0143034316666413

Sinatra, G.M., Broughton, S.H., & Lombardi, D. (2016). Emotions in science education. In R.

Pekrun and L. Linnenbrink-Garcia (Eds.). *International handbook of emotions in education* (pp. 415-436). Routledge.

Skinner, E.A., & Chi, U. (2012). Intrinsic motivation and engagement as "active ingredients" in garden-based education: Examining models and measures derived from self-determination theory. *The Journal of Environmental Education, 43*(1), 16–36. https://doi.org/10.1080/00958964.2011.596856

Skinner, E.A., Furrer, C., Marchand, G., & Kinderman, T. (2008). Engagement and disaffection in the classroom: Part of a larger motivational dynamic? *Journal of Educational Psychology, 100*(4), 765–781. https://doi.org/10.1037/a0012840

Skinner, E. A., Kindermann, T. A., & Furrer, C. J. (2009). A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educational and Psychological Measurement, 69*(3), 493–525. https://doi.org/10.1177/0013164408323233

Sprinkle, R., Hunt, S., Simonds, C., & Comadena, M. (2006). Fear in the classroom: An examination of teachers' use of fear appeals and students learning outcomes. *Communication Education, 55*(4), 389–402, https://doi.org/10.1080/03634520600879170

Swann, W., Chang-Schneider, C., & McClarty, K. (2007). Do people's self-views matter? Self-concept and self-esteem in everyday life. *American Psychologist, 62*(2), 84–94. https://doi.org/10.1037/0003-066X.62.2.84

Symes, W., & Putwain D.W. (2016). The role of attainment value, academic self-efficacy and message frame in the appraisal of value-promoting messages. *British Journal of Educational Psychology, 86*(3), 446–460. https://doi.org/10.1111/bjep.12117

Symes, W., Putwain D.W., & Remedios, R. (2015). Academic buoyancy and the appraisal of fear appeals, used prior to a high-stakes examination, as threatening or challenging. *School Psychology International, 36*(6), 605–619. https://doi.org/10.1177/0143034315610622

Teimouri, Y., Goetze, J., & Plonsky, L. (2019). Second language anxiety and achievement: a meta-analysis. *Studies in Second Language Acquisition, 41*(2), 363-387. https://doi.org/10.1017/S0272263118000311

Tiedens, L. Z., & Linton, S. (2001). Judgment under emotional certainty and uncertainty: The effects of specific emotions on information processing. *Journal of Personality and Social Psychology, 81*(6), 973–988. https://doi.org/10.1037/0022-3514.81.6.973

Travis, J., A. Kaszycki, M. Geden, and J. Bunde. (2020). Some stress is good stress: The challenge-hindrance framework, academic self-efficacy, and academic outcomes. *Journal of Educational Psychology 112*(8), 1632–1643. http://dx.doi.org/10.1037/edu0000478

Uphill, M. A., Rossato, C., Swain, J., & O'Driscoll, J. M. (2019). Challenge and threat: A critical review of the literature and an alternative conceptualisation. *Frontiers in Psychology, 10*, 1255. https://doi.org/10.3389/fpsyg.2019.01255

Urhahne, D., Chao, S., Florineth, M.L., Luttenberger, S., & Paechter, M. (2011) Academic self-concept, learning motivation and test anxiety of the underestimated student. *British Journal of Educational Psychology, 81*(1), 161-177. https://doi.org/10.1348/000709910X504500

Vally, Z. (2020). Public perceptions, anxiety and the perceived efficacy of health-protective behaviors to mitigate the spread of the SARS-Cov-2/COVID-19 pandemic. *Public Health, 187*, 67–73. https://doi.org/10.1016/j.puhe.2020.08.002

Vasalampi, K., Muotka, J., Malmberg, L-E., Aunola, K., & Lerkkanen, K-M. (2021). Intra-individual dynamics of lesson-specific engagement: Lagged and cross-lagged effects from one lesson to the next. *British Journal of Educational Psychology, 91*(3), 997-1014. https://doi.org/10.1111/bjep.12404

Wang, M., & Holcombe, R. (2010). Adolescents' perceptions of school environment, engagement, and academic achievement in middle school. *American Educational Research Journal, 47*(3), 633–662. https://doi.org/10.3102/0002831209361209

Wentzel, K. R., Muenks, K., McNeish, D., & Russell, S. (2017). Peer and teacher supports in relation to motivation and effort: A multi-level study. *Contemporary Educational Psychology, 49*, 32-45. http://dx.doi.org/10.1016/j.cedpsych.2016.11.002.

Widaman, K.F., Ferrer, E., & Conger, R.D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives, 4*(1), 10-18. https://doi.org/10.1111/j.1750-8606.2009.00110.x

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance use research* (pp. 281–324). American Psychological Association. https://doi.org/10.1037/10222-009

Wigfield, A., & Eccles, J. S. (2020). 35 years of research on students' subjective task values and motivation: A look back and a look forward. In A. J. Elliot (Ed.), *Advances in motivation science* (Vol. 7, pp. 161–198). Academic Press.

Wigfield, A., Tonks, S., & Klauda, S. L. (2016). Expectancy-value theory. In K.R. Wentzel and D. B. Miele (Eds.) *Handbook of motivation at school* (pp. 55-74). Routledge.

Table 1

Descriptive Statistics for Study Variables

	Mean	SD	ω	ρ_I	Skewness	Kurtosis	Factor Loadings $_W$	Factor Loadings $_B$
T_1 Fear Appeal Frequency	2.63	1.28	.80	.22	0.30	-0.88	.60 - .84	.82 - .92
T_1 Challenge Evaluation	3.37	0.97	.80	.03	-0.32	-0.91	.62 - .75	.59 - .85
T_1 Threat Evaluation	2.62	1.21	.88	.11	-0.29	-0.85	.72 - .81	.75 - .92
T_1 Behavioral Engagement	3.94	0.87	.87	.06	-0.67	0.49	.70 - .74	.83 - .89
T_2 Fear Appeal Frequency	2.85	1.12	.81	.25	0.10	-0.91	.56 - .81	.87 - .94
T_2 Challenge Evaluation	3.38	0.82	.81	.04	-0.32	0.01	.65 - .78	.69 - .84
T_2 Threat Evaluation	2.56	1.04	.89	.14	0.28	-0.68	.73 - .82	.72 - .92
T_2 Behavioral Engagement	3.97	0.84	.90	.07	-0.86	1.82	.72 - .78	.73 - .96
Year 9 Mathematics Grade	3.46	1.65	—	.29	0.39	-0.05	—	—
GCSE Mathematics Grade	4.85	1.98	—	.28	0.15	-0.47	—	—

Table 2

Latent Bivariate Correlations for Study Variables

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. T_1 Fear Appeal Frequency	—	.23***	.19***	.02	.34***	.02	.08	.06	-.05	-.06	-.07*	-.01	.02
2. T_1 Challenge Evaluation	-.19	—	-.32***	.47***	.03	.51***	-.32***	.31***	.10**	.21***	-.09**	-.06	-.03
3. T_1 Threat Evaluation	.33*	-.60***	—	-.18***	.08	-.23***	.59***	-.07	-.16**	-.31***	.25***	.05	.07
4. T_1 Behavioral Engagement	-.18	.75***	-.41***	—	.06	.49***	-.07	.48***	.18***	.29***	.07*	-.01	-.01
5. T_2 Fear Appeal Frequency	.35***	-.15	.25*	-.14	—	.13*	.19***	.07	-.06	-.08	-.08*	-.01	-.06
6. T_2 Challenge Evaluation	.06	.54***	-.11	.53***	.02	—	-.35***	.49***	.09*	.22***	-.08*	-.04	-.05
7. T_2 Threat Evaluation	.29*	-.49***	.68***	-.33**	.44***	-.41***	—	-.21***	-.18***	-.33***	.23***	.03	.02
8. T_2 Behavioral Engagement	.11	.29*	-.12	.30*	.06	.60***	-.20	—	.10**	.26***	.07	.01	-.02
9. Year 9 Mathematics Grade	-.36***	.69***	-.74***	.63***	-.40***	.28	-.68***	.25	—	.53***	.06	.04	-.04
10. GCSE Mathematics Grade	-.38***	.60***	-.75***	.58***	-.38***	.18	-.70***	.14	.81***	—	-.02	-.03	-.12***
11. Gender	—	—	—	—	—	—	—	—	—	—	—	—	—
12. Age	—	—	—	—	—	—	—	—	—	—	—	—	—
13. Free School Meals	—	—	—	—	—	—	—	—	—	—	—	—	—

Note. Correlation coefficients from the within-level portion of the model above, and from the between-level portion of the model below, the diagonal. Gender, age, and free school means were included only as within-level variables.

* p <. 05. ** p <.01. *** p <.001.

Table 3

Standardized Regression Coefficients from the ML-SEM

	Y9M	T_1 FA	T_1 CH	T_1 TH	T_1 BE	T_2 FA	T_2 CH	T_2 TH	T_2 BE	GCSE
Within-Level Coefficients:										
Year 9 Mathematics Grade (Y9M)		-.05 (.03)	.10 (.03)	-.17 (.05)	.18 (.03)	-.03 (.04)	.03 (.04)	-.10 (.04)	.01 (.04)	.42 (.07)
T_1 Fear Appeal Frequency (T_1 FA)						.33 (.04)	.11 (.04)	-.02 (.04)	.02 (.05)	
T_1 Challenge Evaluation (T_1 CH)							.52 (.04)		.14 (.05)	
T_1 Threat Evaluation (T_1 TH)								.56 (.04)	.02 (.04)	
T_1 Behavioral Engagement (T_1 BE)									.39 (.05)	
T_2 Fear Appeal Frequency (T_2 FA)										-.05 (.04)
T_2 Challenge Evaluation (T_2 CH)										.03 (.05)
T_2 Threat Evaluation (T_2 TH)										-.21 (.05)
T_2 Behavioral Engagement (T_2 BE)										.20 (.05)
GCSE Mathematics Grade (GCSE)										
Gender	.05 (.03)	-.07 (.03)	-.10 (.03)	.26 (.03)	.06 (.03)	-.05 (.03)	-.05 (.03)	.10 (.03)	.05 (.03)	-.01 (.03)
Age	.04 (.03)	-.02 (.03)	-.07 (.04)	.05 (.03)	-.02 (.03)	-.01 (.03)	-.02 (.03)	.01 (.03)	.01 (.03)	-.05 (.03)
Free School Meals	-.04 (.03)	.02 (.03)	-.02 (.03)	.06 (.03)	.01 (.03)	-.06 (.04)	-.05 (.03)	-.01 (.03)	-.02 (.03)	-.09 (.03)
Between-Level Coefficients:										
Year 9 Mathematics Grade (Y9M)		-.28 (.14)	.36 (.14)	-.63 (.07)	.51 (.09)	-.22 (.10)	.18 (.18)	-.09 (.13)	.20 (.19)	.76 (.11)
T_1 Fear Appeal Frequency (T_1 FA)						.28 (.13)	.13 (.17)	.05 (.11)	.05 (.18)	
T_1 Challenge Evaluation (T_1 CH)							.60 (.12)		.18 (.31)	
T_1 Threat Evaluation (T_1 TH)								.71 (.12)	-.01 (.21)	
T_1 Behavioral Engagement (T_1 BE)									.04 (.26)	
T_2 Fear Appeal Frequency (T_2 FA)										.05 (.07)
T_2 Challenge Evaluation (T_2 CH)										.04 (.10)
T_2 Threat Evaluation (T_2 TH)										-.34 (.12)
T_2 Behavioral Engagement (T_2 BE)										-.10 (.08)

Note. Standard errors in parentheses.

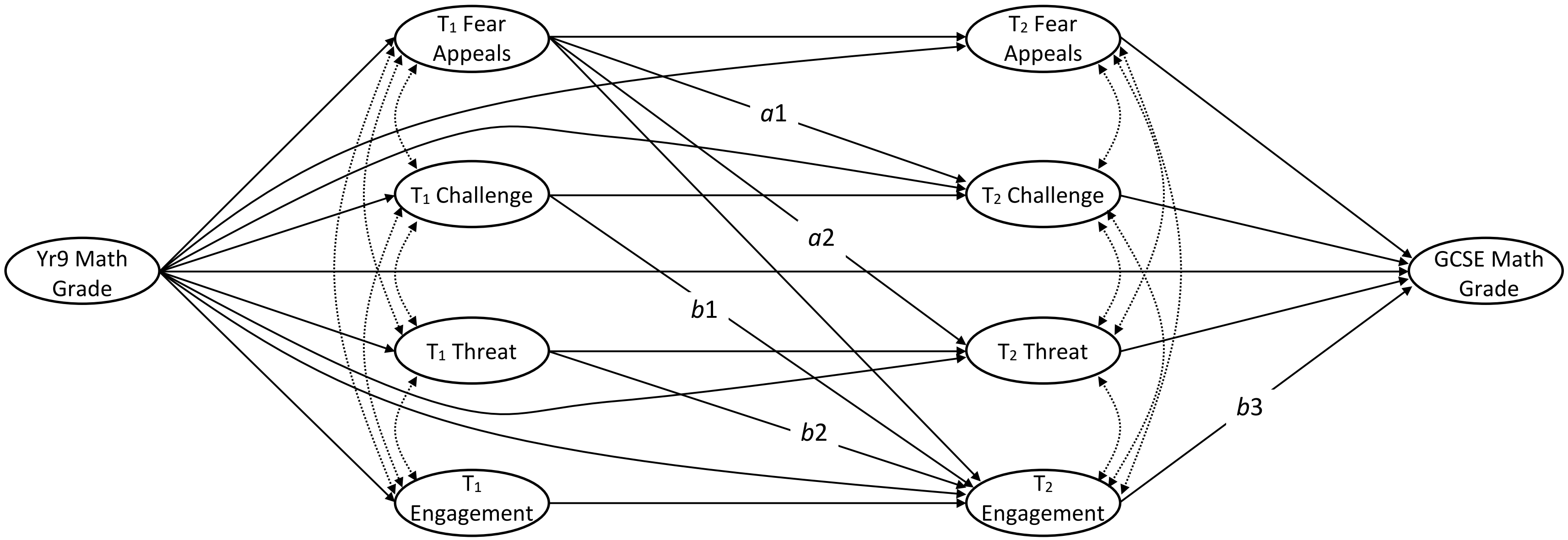
Table 4

Indirect Paths from the ML-SEM

Indirect Paths	β	*SE*	95% CIs
Within-Level Portion of the ML-SEM:			
Fear Appeals to Behavioral Engagement			
Via Challenge Evaluation	.020	.010	.001, .037
Via Threat Evaluation	-.001	.001	-.003, .002
Challenge Evaluation to GCSE Mathematics Grade			
Via T_2 Behavioral Engagement	.035	.015	.011, .059
Fear Appeals to GCSE Mathematics Grade			
Via T_2 Challenge Evaluation and Behavioral Engagement	.004	.002	-.008, .000
Year 9 Mathematics Grade to GCSE Mathematics Grade			
Via T_2 Threat Evaluation	.022	.009	.007, .036
Via T_1 and T_2 Threat Evaluation	.020	.006	.010, .030
Via T_1 and T_2 Behavioral Engagement	.014	.005	.006, .021
Via T_1 Challenge Evaluation and T_2 Behavioral Engagement	.003	.001	.000, .005
Between-Level Portion of the ML-SEM:			
Year 9 Mathematics Grade to GCSE Mathematics Grade			
Via T_1 and T_2 Threat Evaluation	.152	.058	.056, .247

Figure 1

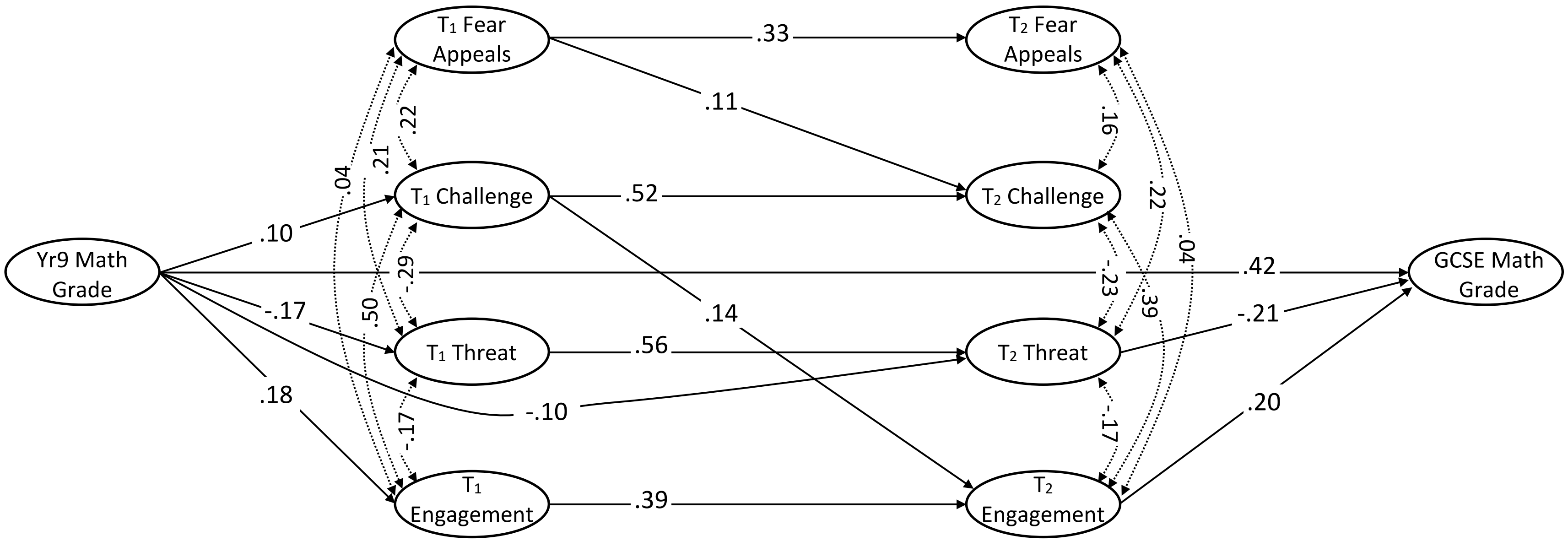
Hypothesized Model



Note. Solid lines represent structural paths and dotted lines correlations.

Figure 2

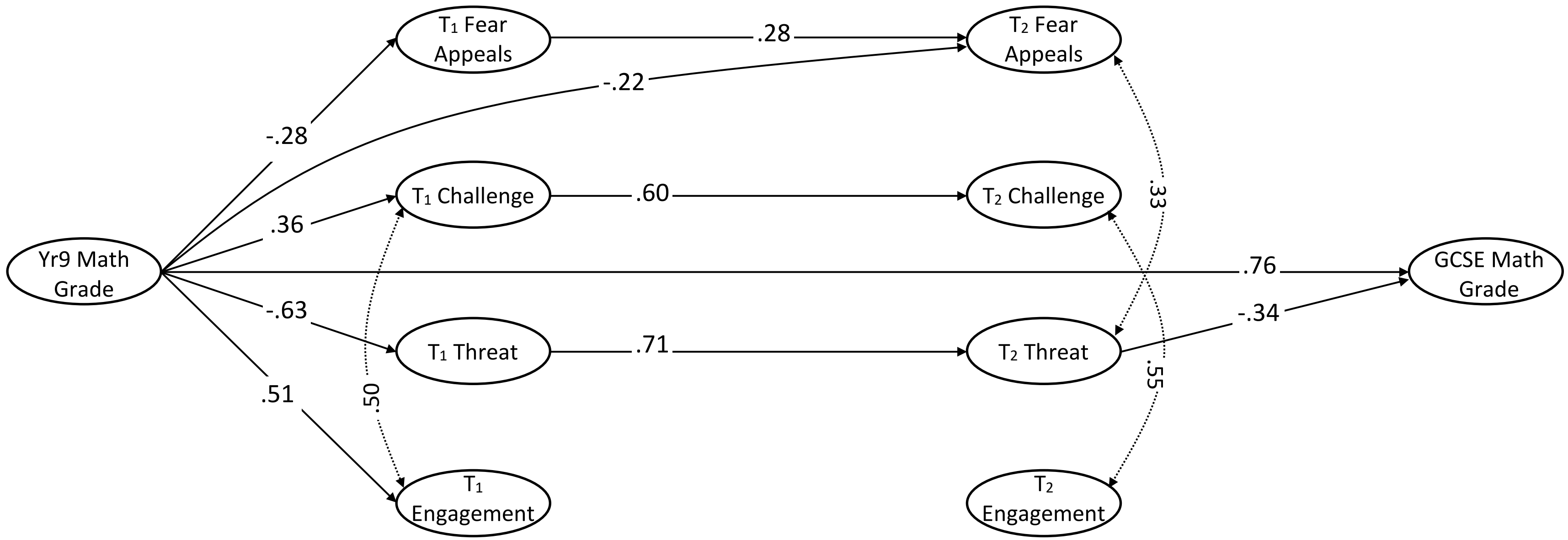
Statistically Significant Standardized Coefficients from the Within-Level Portion of the ML-SEM



Note. Solid lines represent structural paths and dotted lines correlations. Paths for gender, age, and FSM, were omitted for simplicity

Figure 3

Statistically Significant Standardized Coefficients from the Between-Level Portion of the ML-SEM



Note. Solid lines represent structural paths and dotted lines correlations. Paths for gender, age, and FSM, were omitted for simplicity

Warning Students of the Consequences of Examination Failure: An Effective Strategy for Promoting Student Engagement?

- Supplementary Materials -

This document contains materials designed to supplement the main text. The materials include the following:

Detailed Analytic Plan

Missing Data Analysis

Descriptive Statistics

Latent Bivariate Correlations

Longitudinal Measurement Invariance

Tables S1 and S2

Detailed Analytic Plan

First, in a preliminary analysis, we examined basic pre-requisites for multi-level modelling by checking the variance components of variables (i.e., proportion of data present at the individual and class levels) using the ICC_1 (ρ_I) statistic and the consistency of fear appeal frequency (a class climate variable) using the ICC_2 statistic. In addition, assumptions regarding missing data were checked. Second, a measurement model comprising fear appeal frequency and evaluation, and behavioral engagement, was tested using a multi-level confirmatory factor analysis (ML-CFA). Between-level indicators were then class-aggregated from grand mean-centered participant responses and within-level indicators were class mean-centered. Year 9 and GCSE mathematics achievement were added to the measurement model (also in a class-aggregated and class mean-centered form for the between- and within-level indicators respectively) and sociodemographic covariates (gender, age, and FSM) added to the within-level portion of the model. Latent bivariate correlations were estimated in the ML-CFA. Third, longitudinal measurement invariance for fear appeal frequency and evaluation, and behavioral engagement, across T_1 and T_2 waves of data collection was tested using a series of ML-CFAs. Fourth, and finally, a multi-level structural equation model (ML-SEM) was used to test hypotheses.

All models were estimated using the M*plus* v.8 software (Muthén & Muthén, 2017). Goodness-of-fit for latent models were assessed using the root mean error of approximation (RMSEA), standardized root mean residual (SRMR), confirmatory fit index (CFI), and Tucker-Lewis index (TLI). Based on simulation studies, Hu and Bentler (1999) recommend RMSEA ≈ .06, SRMR ≈ .08, and CFI/TLI ≈ .95, for a good model fit. Various methodologists, however, have cautioned against too strict an application of these criteria when assessing complex models based on naturalistic data (e.g., Heene et al., 2011; Lance et al., 2006). In addition, these widely cited criteria are based on single level models.

Multi-level models estimated in M*plus* generate RMSEA, CFI, and TLI indices for the entire model and do not differentiate between the within- and between-level portions. As multi-level models typically have larger sample sizes at the individual level, the RMSEA, CFI, and TLI indices will likely be more sensitive to the within-level portion of the model (Hsu et al., 2015). Although separate SRMR indices for the within- (SRMRw) and between- ($SRMR_B$) level portions of a multi-level model are estimated in M*plus*, studies are only beginning to assess the sensitivity of these indices to misspecification. Based on simulated data, Pagett and Morgan (2021) cautiously advise an $SRMR_B$ of ≤.08 for a good fitting model when the number of L2 units is >100, estimated using maximum likelihood with robust standard errors. In their research, SRMRw and $SRMR_B$ indices did not always reliably differentiate between correctly and incorrectly specified models under suboptimal conditions (especially when the number of L2 units was ≤ 100) and high $SRMR_B$ indices were found even when models were correctly identified. Accordingly, $SRMR_B$ indices of >.08 should be interpreted only as weak potential evidence for misspecification. These findings further underscore the importance of using model fit indices in conjunction with other sources of information (e.g., factor loadings, modification indices, residual variances, alternative fitting models) in order to assess misspecification.

Missing Data Analyses

In common with many longitudinal studies, there was substantial attrition (32.5%) resulting from the combination of students being absent from school or choosing not to participate. Little's omnibus test for data missing completely at random (MCAR; Little, 1988) was statistically significant ($p < .001$) indicating some systematic variation in missingness. Following the recommendations of Nicholson et al. (2017), to identify the cause(s) of the missingness, T_2 data were re-coded as being absent or present. A series of *t*-tests (for continuous variables) and logistic regressions (for gender and FSM) were conducted to predict missingness in T_2 data from T_1 variable scores. Participants eligible for FSM were more likely to have missing data for T_2 fear appeal frequency ($B = .31$, $p = .02$), T_2 challenge evaluation ($B = .31$, $p = .02$), T_2 threat evaluation, ($B = .32$, $p = .03$), and T_2 behavioral engagement ($B = .33$, $p = .01$). All other tests for missingness were not statistically significant (*p*s >.05).

When the sources of missing data can be identified, as we have done here, data can be treated under the missing at random (MAR) assumption (Nicholson et al., 2017). We utilized Full Information Maximum Likelihood (FIML) in subsequent analyses to manage missing data. This powerful, model-based algorithmic approach has been shown to produce unbiased estimates under the MAR assumption, when the variables responsible for the missingness are included in models (Nicholson et al., 2017). Furthermore, FIML has shown unbiased estimates when missing data is substantial (Enders, 2010), as is the case in the present study, and to be a robust method of handling missing data in longitudinal studies (Jeličič et al., 2009).

Descriptive Statistics

With one exception, data showed skewness and kurtosis values within ±1. The exception was T_2 behavioral engagement that showed a leptokurtic distribution. To account for this slight deviation in normality, subsequent latent variable analyses used the maximum likelihood estimator with robust standard errors. All variables showed good internal consistency (McDonald's $\omega \geq .79$). The proportion of variance that was attributable to the class level (the ICC_1 statistic or ρ_I) was substantial for fear appeal frequency (ρ_Is = .22 and .30 for T_1 and T_2, respectively) and mathematics grade (ρ_Is = .28 and .29, for GCSE and Year 9 grades, respectively). Although somewhat smaller, there was still a relatively large proportion of class-level variance for threat evaluation (ρ_Is = .09 to .14). The proportion of class-level variance for challenge evaluation was smaller (ρ_Is = .01 to .03).
Given that class climate variables typically show ρ_Is <.1, and rarely > .3 (e.g., Bliese, 2000; Marsh et al., 2008), the ρ_I statistics (especially those for fear appeal frequency) warranted a multi-level approach to modelling data. The ICC_2 statistic provides an estimate of the reliability of aggregated class climate variables built from aggregated student responses; ICC_2 >.7 is considered as acceptable (Lee, 2000). In the present study, the ICC_2 statistics were .80 and .86 for fear appeal frequency at T_1 and T_2, respectively.

Latent Bivariate Correlations

To estimate latent bivariate correlations, we added gender (0 = male, 1 = female; participants who responded with "other" or "prefer not to say" were coded as missing data), age, and FSM (0 = not eligible, 1 = eligible) to the within-level portion of the model as manifest variables. Year 9 and GCSE mathematics grades were added to the model as single-item indicators. To account for the likelihood of some degree of measurement error in the assessment of mathematics grade, the factor loading was fixed to $\lambda = 1$ and the corresponding residual variance (σ_ε) adjusted for reliability (ρ). The variance for mathematics grades (Year 9 = 1.648; GCSE = 1.982) was multiplied by 1- ρ (Brown, 2006; Little, 2013). We used the lower value ($\rho = 0.74$) from existing reliability estimates (Bramley & Dhawan, 2010; Dhawan & Bramley, 2012) as a conservative value. The ML-CFA showed an acceptable fit to the data, $\chi^2(539) = 1052.03$, $p < .001$, RMSEA = .026, $SRMR_W = .032$, $SRMR_B = .085$, CFI = .951, TLI = .933.

Longitudinal Measurement Invariance

A precondition for the longitudinal modelling of data is measurement invariance; the measurement properties of constructs (namely, fear appeals and their evaluation, and behavioral engagement) must be equivalent at both waves of data collection (Widaman & Reise, 1997). We followed Merideth's (1993) approach of initially testing a configural model where all parameters at T_1 and T_2 were freely estimated. This was followed by testing a model of weak invariance where item factor loadings at T_1 and T_2 were constrained to be equal (metric invariance), a model of strong invariance where item intercepts at T_1 and T_2 were constrained to be equal (scalar invariance), and finally a model of strict invariance where item residual invariances at T_1 and T_2 were constrained to be equal (residual invariance).

Each successively constrained model was compared to the previous using the change in RMSEA, CFI, and TLI fit indices. Simulation studies have suggested that equivalence is shown by ΔRMSEA of <.015 and ΔCFI and ΔTLI indices of <.01 (Chen, 2007; Cheung & Rensvold, 2002). Results of invariance tests are reported in Table S2. Fear appeals and their evaluation showed strict measurement invariance. Behavioral engagement did not demonstrate scalar invariance (ΔTLI = -.014). Accordingly, equality constraints for item intercepts were relaxed one at a time to identify the source of the invariance. Behavioral engagement demonstrated partial scalar invariance when the intercept for one between-level item was freed. Partial scalar invariance is sufficient for the longitudinal modeling of constructs (Widaman et al., 2010).

References

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multi-level theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). Jossey-Bass.

Bramley, T. & Dhawan, V. (2010). *Estimates of reliability of qualifications*. The Office of Qualifications and Examinations Regulation.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. The Guilford Press.

Chen, F.F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Structural Equation Modeling 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Dhawan, V. & Bramley, T. (2012). *Estimation of inter-rater reliability*. The Office of Qualifications and Examinations Regulation.

Enders, C. K. (2010). *Applied missing data analysis*. Guilford.

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: a cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, *16*(3), 319-336. https://doi.org/10.1037/a0024917

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jeličič, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology, 45*(4), 1195-1199. https://doi.org/10.1037/a0015665

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria what did they really say? *Organizational Research Methods*, *9*(2), 202–220. https://doi.org/10.1007/s10984-006-9015-710.1177/1094428105284919

Lee, V. (2000). Using hierarchical linear modelling to study social contexts: The case of school effects. *Educational Psychologist, 35*(2), 125-141. https://doi.org/10.1207/S15326985EP3502_6

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*, 1198–1202. https://doi.org/10.2307/2290157

Little, T.D. (2013). *Longitudinal structural equation modelling*. The Guilford Press.

Marsh, H.W., Martin, A. J., & Cheng, J. H. S. (2008). A multi-level perspective on gender in classroom motivation and climate: Potential benefits of male teachers for boys? *Journal of Educational Psychology*, *100*(1), 78–95. https://doi.org/10.1037/0022-0663.100.1.78

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525–543. https://doi.org/10.1007/BF02294825

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Nicholson, J. S., Deboeck, P. R., & Howard, W. (2017). Attrition in developmental psychology: A review of modern missing data reporting and practices. *International Journal of Behavioral Development, 41*(1), 143-153. http://dx.doi.org/10.1177/0165025415618275

Widaman, K.F., Ferrer, E., & Conger, R.D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child*

Development Perspectives, 4(1), 10-18. https://doi.org/10.1111/j.1750-8606.2009.00110.x

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance use research* (pp. 281–324). American Psychological Association. https://doi.org/10.1037/10222-009

Table S1

Participant Characteristics for the Two Waves of Self-Reported Data Collection

	T_1		T_2	
	n	%	*n*	%
Gender:				
Male	697	45.6	458	44.4
Female	776	50.7	544	52.7
Other	21	1.4	13	1.3
Prefer not to say	29	1.9	16	1.6
Missing	7	0.5	1	0.1
Ethnic Heritage:				
Asian	121	7.9	89	8.7
Black	26	1.7	15	1.5
White	1257	82.2	850	82.3
Dual Heritage	15	0.9	31	3.0
Other	40	2.6	22	2.1
Prefer not to say	43	2.8	14	1.4
Missing	28	1.8	11	1.1
Total	1,530		1,032	

Table S2
Tests of Measurement Invariance

	χ^2(df)	RMSEA	$SRMR_W$	$SRMR_B$	CFI	TLI	Δ RMSEA	ΔCFI	ΔTLI
Fear Appeal Frequency and Evaluation:									
Configural	478.80 (222)	.028	.034	.084	.962	.948			
Metric Invariance	510.89 (234)	.029	.035	.085	.959	.946	+.001	-.003	-.002
Scalar Invariance	549.64 (252)	.029	.035	.085	.955	.946	.000	-.004	.000
Residual Invariance	587.74 (270)	.028	.035	.085	.952	.946	-.001	-.003	.000
Behavioral Engagement:									
Configural	13.84 (10)	.016	.015	.045	.998	.993			
Metric Invariance	18.04 (14)	.014	.015	.045	.997	.995	-.002	-.001	+.002
Scalar Invariance	39.82 (20)	.026	.015	.048	.988	.981	+.012	-.009	-.014
Partial Scalar Invariance[a]	28.71 (18)	.020	.015	.045	.993	.989	-.006	+.005	+.008

Note. χ^2 statistic p <.001 for all fear appeal frequency and evaluation models and p >.05 for all behavioral engagement models with the exception of the scalar invariance model.
[a]Equality constraint relaxed for one item in the between-level portion of the model ("I participate in the activities and tasks in my GCSE maths class").