REVIEW ARTICLE

Prediction Versus Explanation in Educational Psychology: a Cross-Theoretical Approach to Using Teacher Behaviour to Predict Student Engagement in Physical Education

Michael Noetel · Philip Parker · Theresa Dicke · Mark R. Beauchamp · Nikos Ntoumanis, · Ryan M. Hulteen, et al. *[full author details at the end of the article]*

Abstract

Educational psychology usually focuses on explaining phenomena. As a result, researchers seldom explore how well their models predict the outcomes they care about using best-practice approaches to predictive statistics. In this paper, we focus less on explanation and more on prediction, showing how both are important for advancing the field. We apply predictive models to the role of teachers on student engagement, i.e. the thoughts, attitudes, and behaviours, that translate motivation into progress. We integrate the suggestions from four prominent motivational theories (self-determination theory, achievement goal theory, growth mindset theory, and transformational leadership theory), and aim to identify those most critical behaviours for predicting changes in students' engagement in physical education. Students ($N=1324$ all from year 7, 52% girls) from 17 low socio-economic status schools rated their teacher's demonstration of 71 behaviours in the middle of the school year. We also assessed students' engagement at the beginning and end of the year. We trained elastic-net regression models on 70% of the data and then assessed their predictive validity on the held-out data (30%). The models showed that teacher behaviours predicted 4.39% of the variance in students' change in engagement. Some behaviours that were most consistently associated with a positive change in engagement were being good role models ($\beta=0.046$), taking interest in students' lives outside of class ($\beta=0.033$), and allowing students to make choices ($\beta=0.029$). The influential behaviours did not neatly fit within any single motivational theory. These findings support arguments for integrating different theoretical approaches, and suggest practitioners may want to consider multiple theories when designing interventions. More generally, we argue that researchers in educational psychology should more frequently test how well their models not just explain, but *predict* the outcomes they care about.

Engagement is a multidimensional process that drives student learning (Reeve & Lee, 2014; Skinner & Belmont, 1993). It is a critical pathway that translates motivation into learning (Reeve et al., 2019). As a result, many motivational theories aim to account for how teachers influence engagement (Lazowski & Hulleman, 2016). To promote engagement, teachers may be expected to satisfy basic psychological needs (self-determination theory; Ryan & Deci, 2020), to be transformational leaders (transformational leadership theory; Bass & Riggio, 2006), to promote a mastery climate (achievement goal theory; Huang, 2012), and to build growth mindsets (growth mindset theory; Blackwell et al., 2007). While theoretical work is ongoing, one way to help teachers improve engagement now is to explore ways we can best predict engagement using multiple current theories. By focusing less on how particular models *explain* changes in student engagement, we can focus more on the specific behaviours that best *predict* change in student engagement. It is this latter approach we take here.

Theories provide cohesive explanations of how teacher behaviour influences student motivation and engagement. For example, Dweck (2008) argues that effort-oriented praise leads students to develop beliefs that their abilities are incrementally improved through hard work (a growth mindset), which makes student motivation robust to inevitable failures. Explanatory theories, like growth mindsets, are critical for the advancement of educational psychology, but parsimonious explanations can lead to poor predictions (Shmueli, 2010). Meta-analyses have shown that a growth mindset might be influential over time (Funder & Ozer, 2019) but only modestly predicts achievement ($r=0.10$; Sisk et al., 2018). As described in other articles of this special issue (e.g. Dinsmore et al., n.d.), we may need to combine multiple theories to sufficiently explain and predict educational psychology phenomena. Drawing on behaviours identified in multiple theories can increase our ability to predict outcomes. This is because we are not constrained by adherence to a particular set of mechanisms but are focused on what predicts best regardless of the source of the predictors (Shmueli, 2010).

In addition, theories overlap. For example, all four theories mentioned above emphasise the importance of improvement-oriented specific feedback. It is possible that the conceptual differences between theories are exaggerated, leading to 'empirical redundancy' (Le et al., 2010) or jingle-jangle problems (Marsh et al., 2019; Reschly & Christenson, 2012). Empirical redundancy is where separate constructs are so tightly correlated that they effectively describe the same underlying construct (Le et al., 2010). Somewhat ironically, jingle-jangle problems describe the same essential phenomena: where researchers use a single term for different things (jingle) or different terms for the same thing (jangle; Reschly & Christenson, 2012). These problems frequently arise in psychology literature on motivation and engagement (Hoch et al., 2018; Le et al., 2010; Marsh et al., 2019; Reschly & Christenson, 2012). For example, Marsh et al. (2019) found math

self-concept, generalised self-efficacy, and outcome expectancies to be empirically indistinguishable. Cross-theoretical approaches can manage this overlap by simultaneously assessing the predictions of each theory (Le et al., 2010; Marsh et al., 2019; Reschly & Christenson, 2012).

In this paper, we report findings from a cross-theoretical analysis with a view toward designing interventions that improve student outcomes. Interventions to improve motivation and engagement often draw from multiple theories (Lazowski & Hulleman, 2016), but the specific behaviours chosen from each theory can be arbitrary rather than informed by data. Theories are useful for intervention design because they hypothesise behaviours likely to increase engagement via an established causal model, but interventions usually group together a range of behaviours that all purport to influence outcomes (e.g. Beauchamp et al., 2011; Cheon et al., 2012; Digelidis et al., 2003). In this paper, we assess which teacher behaviours from four theories of motivation best predict changes in student engagement for physical education during secondary school, with a particular focus on students with lower socio-economic advantage.

In doing so, we hope to address five of the seven goals of this special issue. We assess the points of convergence, divergence, and creative synthesis that comes from integrating different theoretical models (goal 1). Specifically, we describe the behavioural recommendations that overlap between four well-researched theories of engagement, we show which behaviours are unique to those theories, and we assess which behaviours are most important when simultaneously accounting for the predictions of all four theories. In turn, we also assess gaps that are apparent across each theory (goal 2); namely, if any of our four 'competing' theories are weak predictors of student engagement, it may suggest that the theory is focused on an ancillary (rather than core) driver of engagement. By simultaneously assessing the behaviours prescribed by four theories, we assess competing hypotheses between the different theories around which behaviours should be influential (goal 4), and we assess the predictive validity of statistical models that simultaneously leverage multiple perspectives on improving engagement (goal 5). That is, we test the combined explanatory power of integrating these psychological theories, while also assessing whether the predictions of some theories are more powerful than others. In order to achieve these goals, we explore a different fundamental epistemological basis for our approach (goal 6) by focusing less on explanation and more on prediction. As we will argue, both predictive and explanatory statistics provide useful information about the nature of the world, but the predictive methods are largely neglected in educational psychology. We review the strengths and weaknesses of the predictive approach, and show how it can serve the goals of this special issue by empirically integrating multiple theories in an interpretable manner.

The Differences Between Prediction and Explanation

As noted in other articles of this special issue, the key goals of educational psychology are to explain and predict useful educational outcomes (Dinsmore et al., n.d.). Both of these are important goals: without explanation, we cannot describe the

mechanisms underpinning a phenomena; without prediction, we cannot assume our models will continue to work in the future. Like most of the social sciences, educational psychology has focused on the former but has seldom explicitly used tools focused on the latter (Breiman, 2001; Parker et al., 2018; Scheinost et al., 2019; Yarkoni & Westfall, 2017). A review of the literature on models derived from education panel data found only a handful of studies that rigorously reported predictive accuracy (Parker et al., 2018). Some studies use methods like cross-validation to estimate predictive accuracy. Cross-validation generally involves repeatedly using the same set of data to both train and test the quality of the models (Hastie et al., 2009; James et al., 2013). For example, fivefold cross-validation cuts the data into 5 random slices, trains the model 5 times, and, each time, tests the fit of the model on one of the slices not used for training.

The key problem with this approach is the word 'repeatedly'. Using cross-validation, only once is a good estimate for generalisability. However, few researchers use cross-validation in this way. They generally use cross-validation for finding the best models that fit the data. For example, in the model we use below (elastic-net regression), researchers must choose values for two 'hyperparameters' that reflect how the models penalise coefficients. Rather than choosing these parameters prospectively, researchers often use cross-validation on the training data to find hyperparameters that best fit the data. In ordinary least squares regressions, researchers might try models with greater or fewer covariates, interactions, or non-linear terms. By repeatedly using the same data for training and testing, the researcher can tweak their models and parameters so it best fits the data. This sounds like an advantage, but it means the resulting model may not generalise well to new data, not used in training (a problem called 'overfitting', described below). Put another way, by learning from the data via cross-validation or other methods (like AIC), then continuing to refine one's model in response, researchers are likely to overfit to the data. We need a metric of success that is not as vulnerable to overfitting.

As a result, the least biased assessment of predictive models involves testing those models on new data, not used to train the model (Hastie et al., 2009; James et al., 2013; Yarkoni & Westfall, 2017). Best practice is that you choose among competing models using cross-validation or AIC or any other metric you feel is valuable and then check it on the new (or unseen) data once. Checking more than once means learning from that data and succumbing to overfitting once again. For example, when seeking to predict, R-squared values should not be calculated on data used in building the model (seen data), but instead calculated using either new data or specially reserved data (unseen data).[1] This is the quintessential difference between explanatory and predictive approaches: where explanatory approaches try to understand mechanisms explaining data in the past, predictive models try to create the

---

[1] There are no hard rules regarding how much data should be used for building models and how much should be reserved for testing models. The more data held out for testing, the more accurately results from the testing process are likely to reflect the general population. However, in a world with finite data, more data for testing means less data available for building models, meaning the models themselves may be poorer fits for the general population (Hastie et al., 2009; James et al., 2013). As a result, a common choice in predictive analytics is a 70% training to 30% testing split.

best model for predicting new data in the future (Breiman, 2001). The ultimate test of these methods involves building a model that makes a set of prospective predictions, then collecting new data and seeing how well those predictions perform. This is arduous, so a robust, pragmatic alternative is to do this retrospectively: to quarantine some data that are never used to build the model and use that data for testing the models being built.

One consequence of this method of evaluation—in contrast with explanatory models—is that researchers may *deliberately reduce* the fit of their models on the training data in order to help those models generalise. For example, in our methods below, we use a penalised regression model that produces fewer and smaller regression coefficients than would optimally fit the training data, with the consequence that the model generalises better to new data (Hastie et al., 2009; James et al., 2013). Explanatory models will seldom make their models try to fit the data 'less well', but as we discuss below, this can mean models in educational psychology risk overfitting, generalising poorly, and failing to replicate (Scheinost et al., 2019; Yarkoni & Westfall, 2017).

This predictive approach is seldom used in education (Parker et al., 2018). Education researchers may not use these approaches because they are unfamiliar with these recommendations (Yarkoni & Westfall, 2017), or they may not focus on prediction because they mistakenly believe that explanation inherently leads to prediction (Shmueli, 2010).

The Culture and Goals of Explanation Versus Prediction

A *complete* explanatory model will provide an excellent prediction. Understanding Newton's laws of motion means we can predict the time of high tide years into the future. However, errors in prediction can point to gaps in the explanatory model. Prediction errors in the perihelion of Mercury led Einstein to develop his new explanatory model: general relativity. As a result, both prediction and explanation are important for advancing science. This paper aims to describe how predictive models can be used to test and complement explanatory approaches in educational psychology. We hope it helps researchers integrate theories and identify gaps in our explanatory models. For the field to advance, it needs to engage in both approaches.

In practice, explanation and prediction reflect two cultures in statistical modelling with different established norms (Breiman, 2001). Explanation approaches aim to build statistical models that replicate specific natural processes. These processes are often derived from a priori theory, are parsimonious, and are clearly articulated. For example, self-determination theory researchers hypothesise that more engaged students are driven by more autonomous forms of motivation, and that motivation is nurtured by satisfying the students' psychological needs, and the teacher's behaviour is instrumental in whether or not those needs are satisfied (Reeve & Cheon, 2021; Ryan & Deci, 2017; Vasconcellos et al., 2019). The goal of explanatory modelling is to provide insight into some form of causal structure of the natural world. Such models typically include inferences to populations (e.g. confidence intervals), are generally simple in nature (e.g. multiple regression), and typically assess goodness

of fit on the same data used to construct the model (i.e. seen data). As a result, standard null-hypothesis significance testing is not 'predictive' because the inferential statistics are almost always retrospective (Scheinost et al., 2019; Yarkoni & Westfall, 2017). They seek to make inferences about some underlying, natural phenomena. However, explanatory models tend to make these inferences by only looking backward at the data used to develop the model. This is generally accepted because the goal is to clearly articulate the processes under investigation—to understand them. Importantly, although such models often include metrics that look like they assess predictive accuracy, like R-squared, these would not be considered valid under a predictive approach (because they are not derived from unseen data).

In contrast, the predictive culture does not prioritise explaining the mechanisms underpinning a phenomenon: it instead optimises directly for accuracy on future data (Breiman, 2001). As a result, the norms allow for more complexity in modelling (Breiman, 2001; Yarkoni & Westfall, 2017). Predictive models treat natural processes as complex and interacting—often too complex to model explicitly. The goal here is to build a model that has ongoing value; that is, as new data comes along (i.e. a new single case of data), the model can continue to make precise predictions. As a result, predictive models allow for more complexity to account for these processes (Breiman, 2001). For example, random forest models, support vector machines, and neural networks are all flexible models that often provide better predictive accuracy than simple linear regression because they can handle hundreds of predictors while looking for interactions and non-linear relationships (Hastie et al., 2009; James et al., 2013). But, those same features can also make causal inference and interpretation more difficult: for example, deep neural networks are notoriously hard to interpret (Christian, 2020). Still, there are many approaches to predictive modelling that allow for interpretable explanations (Molnar, 2021), which we exploit below.

Overfitting and Underfitting

This added flexibility afforded by predictive models makes the predictive culture more attuned to 'overfitting'. As noted above, this is where a statistical model predicts that data used to train the model, but does not generalise to new data (Yarkoni & Westfall, 2017). Overfitting is akin to just remembering the answers on a multiple-choice practice test, but struggling when new questions are on the exam. At the other extreme, models can 'underfit'. Underfitting occurs when the models are too simple to optimally follow the data, akin to the heuristic of 'always answer c' on a multiple choice test. On a practice test (your 'training data'), it is easy to know if your model is underfitting: the predictions will be poor. However, it is impossible to know if models are overfitting, because a complex model might be doing an excellent job of predicting the 'answers'. As mentioned above, the best way to know if a model is overfitting is to see how well it works on a new test (unseen data; Hastie et al., 2009; James et al., 2013). In prediction, the R^2 on unseen data is what matters most because it represents the best estimate of the 'true' predictive accuracy. In contrast to explanatory models, deriving predictive models with a high R^2 on the training data is not always a good thing, because it is *the* key symptom of overfitting.

A model that perfectly predicts the training data ($R^2_{training} = 1$) will almost certainly be overfitting and will struggle on new data. Therefore, researchers aiming to predict will often deliberately reduce the R^2 to ensure the models better predict unseen data. For example, as we do in this paper, researchers may 'penalise' or 'regularise' models so they are less flexible but more likely to predict future data.

In truth, problems of overfitting and underfitting occur both with predictive and explanatory models. In explanatory models, there are dozens of ways of analysing the same dataset (Silberzahn et al., 2018). Researchers have many degrees of freedom regarding how many predictors to include, how many control variables, what models to use, and how to preprocess data (e.g., how to handle outliers or non-normal data). Bad explanatory modelling involves 'throwing the kitchen sink' at the data and publishing the best results. Without prospective registration of all methods, authors can change their models until they best fit their data. Overfitting is one cause of psychology's 'replication crisis' (Yarkoni & Westfall, 2017). Researchers may find a significant result when *just the right number* of control variables are added to their models, but when those same models are tested in other samples, the models fit poorly (Yarkoni & Westfall, 2017). Researchers who focus on prediction often direct more attention to these problems because they transparently try dozens of different, complex models, so more obviously risk overfitting (Yarkoni & Westfall, 2017). They cannot 'throw the kitchen sink' at the data and just use the model that fits the best because this approach is likely to overfit to the training data and perform poorly on the test set. As a result, the habits of the predictive culture—like trying multiple models, tuning them through cross-validation, and testing them on a test set—may be useful for educational psychology. By using these methods, we can test how well our models predict the outcomes we care about, and, in turn, improve our explanatory models.

In this paper, we aim to use predictive models to complement the concerted effort from educational psychologists to *explain* motivation and engagement. Following the suggestions of Yarkoni & Westfall (2017), we are not recommending an abandonment of efforts to explain these phenomena via 'dustbowl empiricism' (Walters, 2016). Instead, we hope to use the strengths of predictive models to eventually inform theory (Shmueli, 2010) and to solve practical challenges in educational psychology. Given a range of compelling theories for how to improve motivation and engagement, each with dozens of behavioural recommendations, what should we focus on when designing an intervention? If we combine these theories, how well can we predict the influence that a teacher will have on a student's engagement?

Engagement in Physical Education Has Wide-Ranging Benefits

Definitions of student engagement typically focus on three elements, including emotion, cognition, and behaviour (Fredricks et al., 2004; Lam et al., 2014). Engagement is particularly important in secondary school, as students typically become less and less engaged (Janosz et al., 2008). As students become less engaged, they are more likely to show disruptive behaviours, struggle academically (Marks, 2000), and drop-out (Archambault et al., 2009), even

after controlling for achievement, socio-economic status, and other covariates (Reschly & Christenson, 2012). Meta-analyses show overall levels of school engagement increase student performance (Lei et al., 2018). Furthermore, students' engagement during secondary school is correlated with their employment success up to 20 years later (Abbott-Chapman et al., 2014). Nurturing student engagement is, thus, critical for their long-term success and flourishing.

In this paper, we focus on engagement in physical education for a number of reasons. Engagement in physical education is important because it has a wide range of possible benefits. Systematic reviews show quality physical education improves cognition (García-Hermoso et al., 2021), academic performance (García-Hermoso et al., 2021), mental health (Andermo et al., 2020), and physical health indicators (Hartwig et al., 2021). However, engagement in physical education declines throughout secondary school (Barkoukis et al., 2010). This means students feel less enjoyment during classes (affective engagement), use fewer problem solving strategies to learn (cognitive engagement), and expend less effort during lessons (behavioural engagement). These declines may put young people's health and development at risk (Bailey, 2006). Children who do not enjoy physical education (low affective engagement) are also less active during their leisure time (Woods et al., 2012). Low cognitive engagement leads to weaker skill acquisition (Kolovelonis et al., 2011), which inhibits long-term physical activity participation (Bremer & Cairney, 2018). Behavioural engagement (e.g. in-class measures of physical activity) is one of the key mechanisms through which physical education leads to health outcomes like cardiorespiratory fitness (Hartwig et al., 2021). As a result, physical education is a useful domain to focus because engagement in the subject is critical for children's health, wellbeing, and academic performance. In addition, physical education has received multidisciplinary attention from education, health, exercise, and sport psychology. As a result, there are a range of diverse theories that propose strategies of improving engagement in physical education, meaning it is a useful domain to focus for a study in this special issue on theoretical integration. In sum, we focus on physical education for this paper because engagement in physical education typically declines across time, those declines have dire long-term consequences, and multiple theories posit competing suggestions for how to best stem that decline.

These declines appear more likely and more damaging for young people with low levels of socio-economic advantage. Socio-economic status (SES) is a reliable predictor of low physical activity (O'Donoghue et al., 2018) and health problems in adults (Marmot, 2005). Children at low SES schools are less likely to be physically active (Peralta et al., 2019), and systematic reviews show SES predicts higher school absenteeism (Sosu et al., 2021) and lower academic performance (Selvitopu & Kaya, 2021). These relationships appear to be mediated by the effects of SES on student engagement (Tomaszewski et al., 2020). So, given the increased risks of disengagement, physical activity, and health for people from low SES backgrounds, we focused our application of predictive models to testing theories of engagement with these students.

Many Theories Describe How Teachers Can Improve Engagement

A range of theories postulate ways in which teachers could stem these declines in motivation and engagement (Lazowski & Hulleman, 2016; Skinner & Pitzer, 2012). For example, teachers could aim to increase students' confidence in their abilities (Dermitzaki et al., 2009), help students feel connected to their teachers and peers (Roorda et al., 2011), and help them see how their learning aligns to their values and goals (Reeve, 2012). Teachers' behaviour strongly influences all of these variables in physical education (Vasconcellos et al., 2019) and across other learning areas (Quin, 2017). Importantly, teachers can learn new motivational behaviours that help improve engagement via these mechanisms (Su & Reeve, 2011; Vasconcellos et al., 2019). For example, providing students with choices and rationales about the activities they complete during a lesson enhances their engagement with the topic (Tsai et al., 2008). As a result, teachers' behaviour is a strong point of leverage for producing positive change in students (Curran & Standage, 2017). In this paper, we draw from four established theories of motivation to assess which behaviours might best improve engagement among students in physical education. While there are many useful, well-supported theories we could have chosen (Lazowski & Hulleman, 2016; Skinner & Pitzer, 2012), we developed our shortlist of theories through informal consultation with researchers, policy makers, and teachers. We chose two frameworks that have been well-researched in physical education: self-determination theory (SDT; Ryan & Deci, 2000) and achievement goal theory (Huang, 2012). We also chose two frameworks that were somewhat less prominent in the physical education literature, but potentially useful for explaining student engagement: implicit theories of ability (Blackwell et al., 2007) and transformational leadership theory (Bass & Riggio, 2006). The theories were selected because they either contained a large number of concrete prescriptions for how teachers might improve engagement (self-determination theory, achievement goal theory, transformational leadership theory) or because teachers and policymakers expressed interest and curiosity in a theory but uncertainty about the relative strength of its evidence (growth mindset theory).

Self-Determination Theory

Self-determination theory hypothesises that humans have three basic psychological needs for competence, autonomy, and relatedness that, when fulfilled, lead to optimal motivation, engagement, and achievement (Ryan & Deci, 2020; Vasconcellos et al., 2019). There are a range of mechanisms underpinning this relationship, explained within the mini-theories within self-determination theory (Ryan & Deci, 2017). For example, Ryan and Deci argue students who experience pressure or psychological control become less driven by intrinsic reasons (cognitive evaluation theory). In contrast, students whose needs are supported by their teachers are more likely to internalise the values and beliefs of their teacher, and pursue learning for their own reasons (organismic integration theory). These reasons can include a desire to feel connected to their teachers and their peers (relatedness motivation theory). As a result, self-determination theory posits that a key driver of student

success is whether their environment supports their psychological needs for competence, autonomy, and relatedness (basic needs theory; Ryan & Deci, 2017).

Teacher behaviours that support these needs include providing clear guidelines and expectations (i.e. competence support), creating a caring environment in the classroom (i.e. relatedness support), and providing opportunities for students to make meaningful decisions about their learning (i.e. autonomy support; Ryan & Deci, 2020). Meta-analyses and systematic reviews have shown that teachers who support basic psychological needs foster motivation and engagement in physical education (Teixeira et al., 2012; Vasconcellos et al., 2019; White et al., 2021). SDT has particular relevance to low SES communities because, compared with their wealthier peers, students in these schools are exposed to more controlling behaviours from their teachers which undermines their perceptions of autonomy.

Achievement Goal Theory

Systematic reviews have also supported the use of achievement goal theory—both the original (Nicholls, 1984) and subsequent revisions of it (Elliot & McGregor, 2001; Elliot et al., 2011)—for promoting outcomes like motivation and engagement in physical education (Jaitner et al., 2019). The theory suggests that students with mastery goals (students who focus on improvement, or succeeding at a task) will be more engaged than those who have performance goals (students who focus on peer comparison; Van Yperen et al., 2014). Similarly, the theory proposes that students who focus on approach goals (e.g. approach-performance goals like winning) are more engaged than those who focus on avoidance goals (e.g. avoidance-performance goals like not losing; Van Yperen et al., 2014). Researchers have argued these goals are important because self-approach goals create a stronger internal locus of control, more reliable feedback that increases self-efficacy, and positively valenced motivation. Self-focused goals are more in a student's control than success against some absolute benchmark (task goals; Elliot et al., 2011) or against other students (other/performance goals; Elliot et al., 2011). Performance goals are particularly risky for students' sense of success because, as Ashley Brilliant said: 'the surest way to remain a winner is to win once, and then not play any more'. Similarly, approach goals are more positively valenced than avoidance goals—it is more 'scary' for a student to fear failure than to want success.

Research in schools supports many of these prepositions. Across subject domains, students with approach goals tend to be more engaged than those with avoidance goals, regardless of whether the goals are mastery or performance oriented (Huang, 2012; Hulleman et al., 2010; Linnenbrink-Garcia et al., 2008; Van Yperen et al., 2014). In physical activity, meta-analyses have shown mastery goals were associated with enjoyment, prosocial behaviour, and engagement (Biddle et al., 2003; Jaitner et al., 2019; Ntoumanis & Biddle, 1999). While the majority of the research has been observational, intervention studies support the causal model where training teachers in building a mastery climate helps students focus on mastery goals, leading to positive student-level outcomes (Jaitner et al., 2019).

Growth Mindset Theory

Growth mindset theory describes a similar causal model to achievement goal theory. As mentioned above, growth mindset theory proposes that students whose teachers foster the belief that ability is non-malleable (i.e. fixed mindset) are less engaged than those who emphasise that ability is incremental and can be improved with effort (growth mindset; Blackwell et al., 2007). The theory argues that these beliefs are implicit, meaning students are seldom explicitly aware of their own mindset (Dweck, 2013). Regardless, Dweck (2008) argues these beliefs are influential because they determine students' agency over their abilities, and therefore influence their self-efficacy and motivation. She argues a growth mindset makes students' self-confidence more resilient to failure, and less vulnerable following success. If a student with a fixed mindset experiences success, they may feel 'smart' or 'talented', but then avoid challenges because doing so increases the threat of failure. Failure could 'show' them they are 'dumb' or 'lack talent', and feel like a strong signal that they would fail with continued effort. Students with growth mindsets may see failure as more reflective of their practice or effort. As a result, these students may be more willing to challenge themselves and persist following failure. Teachers can foster a growth mindset by teaching students about the malleability of ability or by praising students for their effort and improvement, rather than their abilities or talents (Dweck, 2008). Across subject areas, educational interventions to build a growth mindset have, on average, small long-term effects (Sisk et al., 2018). However, in physical activity and sport, meta-analyses have shown that having a growth mindset is reliably connected with adaptive outcomes (e.g. enjoyment, perceived competence, intrinsic motivation) whereas a fixed mindset is associated with maladaptive outcomes (e.g. avoidance goals; Vella et al., 2016).

Transformational Leadership Theory

In organisational psychology, one reliable predictor of adaptive and maladaptive outcomes is the type of leader in charge of the group (Hoch et al., 2018). One well-established model of leadership that is posited to derive better outcomes is that of transformational leadership (Hoch et al., 2018). Transformational leadership occurs when leaders move beyond their own self-interests to empower those being led to achieve higher levels of functioning (Bass & Riggio, 2006). Poor leaders inspire through the use of rewards and punishments (transactional leadership; Bass, 1985), as also hypothesised by self-determination theory (Ryan & Deci, 2020). Transformational leaders, on the other hand, are good role models with high expectations who give followers individualised support to meet intellectual challenges (Bass, 1985). Beauchamp and colleagues adapted transformational leadership theory to the teaching context (Beauchamp et al., 2010, 2014). Aligned with theoretical propositions (Bass, 1985), Beauchamp and colleagues found students were more active at school and during leisure time when they perceived that their teachers displayed four types of behaviours: idealised influence (positive behavioural role modelling),

inspirational motivation (motivating through high expectations), intellectual stimulation (challenging students to examine issues from multiple viewpoints), and individualised consideration (understanding and meeting the needs of individual students). Systematic reviews have found preliminary support for transformational leadership theory in both youth sport (Turnnidge & Côté, 2018) and educational leadership (Gumus et al., 2018). Studies have found that transformational teaching led to higher engagement and motivation in physical education both in primary (Wilson et al., 2012) and secondary settings (Beauchamp & Morton, 2011; Beauchamp et al., 2014). Transformational teachers appear to increase engagement for a range of reasons (Slavich & Zimbardo, 2012). For example, they increase self-efficacy, intrinsic motivation, self-determined motivation, and enjoyment (Beauchamp & Morton, 2011; Wilson et al., 2012), in much the same way as a 'need supportive teacher'. However, for achieving these outcomes, transformational teaching emphasises a range of strategies that are not central to other theories, like building a shared class vision, experiential learning, and being an inspiring role-model. Therefore, transformational teaching seems to operate via similar mechanisms as self-determination theory but with different recommended teaching strategies.

What Teacher Behaviours Best Predict Student Engagement?

As this last example shows, these four theories overlap. For example, they all describe the benefit that students receive when a teacher provides them with clear, personalised learning goals and tailored, task-oriented feedback. It is possible that these behaviours are the strongest predictors of student engagement, which is why each theory considers those behaviours within their hypotheses. If that were true, and we simultaneously assessed the predictive efficacy of each theory, the behaviours common across theories would be most influential. Perhaps those designing interventions should focus on the strategies common to the range of motivational theories. However, each theory also emphasises different mechanisms by which teachers could improve engagement. It is plausible that each theory provides insights into a unique method through which teachers could improve engagement. If that were true, then the best interventions would use the strategies unique to each theory to get the widest coverage of different influences on engagement. Previous research makes this kind of decision-making difficult.

One reason decision-making is difficult is because interventions usually 'bundle' behaviours. For example, when designing their SDT-based interventions for physical education teachers, Cheon et al., (2012; Reeve et al., 2019) focused a range of specific behaviours intended to support psychological needs: taking students' perspective, creating opportunities for meaningful decisions, providing rationales, acknowledging negative feelings, using invitational language, and displaying patience. While these interventions successfully enhanced motivation and engagement (Cheon et al., 2012; Reeve et al., 2019), it is unclear which elements are the active 'ingredients'— the specific strategies from these interventions that are most important for enhancing student engagement in physical education. Additionally, no one theory appears to lead to systematically better interventions for improving motivation. Lazowski &

Hulleman's (2016) meta-analyses of interventions failed to find significant moderation effects by theory. Constraining an intervention to a single theoretical model allows for a good assessment of that theory's causal model, but it could limit the potential benefit of each intervention. For example, an intervention that aims to promote clear goals to increase students' perceived competence, as outlined in SDT, but does not address beliefs about the malleability of ability, might not be as effective as an intervention that targets both objectives. But when researchers create interventions using multiple theories, the interventions are no better at increasing motivation (Lazowski & Hulleman, 2016). Given interventions that use multiple theoretical perspectives demonstrate similar effects to using a single theory (Lazowski & Hulleman, 2016), the method of synthesising different theoretical approaches may be more arbitrary than data-driven. This paper aims to address some of these problems.

One of the few tractable methods for determining which teacher behaviours are most influential is to simultaneously assess the unique influence of the behaviours postulated by the range of theories. In doing so, we aimed to:

- synthesise the predictions from multiple theories (e.g. 'providing students with X increases engagement' where X are the theory-consistent behaviours; special issue goal 1);
- assess both overlapping and competing hypotheses about which teacher behaviours are important (e.g. 'improvement oriented feedback' being common to all theories; 'idealised influence' uniquely emphasised by transformational leadership; goal 4);
- better predict student engagement by using all theories than any single approach (goal 5); and
- provide data-driven recommendations for the specific teacher behaviours that are most influential for engagement (goal 2).

By identifying these most important teacher behaviours, we hoped to provide an empirical basis for future intervention development that could most efficiently promote student engagement.

Accomplishing this goal required a number of steps. First, we aimed to construct a 'mutually exclusive and collectively exhaustive' (Local Burden of Disease WaSH Collaborators, 2020) list of teacher behaviours that influence engagement, as recommended by the four theories. That is, we aimed to create a comprehensive list of teacher behaviours recommended by the four theories (i.e. collectively exhaustive), while grouping behaviours from different theories that overlap (to make the list mutually exclusive). We conducted a systematic process of binning and winnowing (DeWalt et al., 2007; Mâsse et al., 2016) to make a reproducible list of teacher behaviours that have been hypothesised to be influential.

Next, we had a sample of students rate their teachers on the degree to which their teachers used each of those behaviours with them. Finally, we longitudinally assessed which of those behaviours best predicted change in student engagement across the first year of secondary school. We focused on engagement because it balances proximity to important outcomes against the tractability of measuring change within a year. Engagement is a proximal cause of long-term benefits, but it may be

more likely to change within a single school year, compared with distal outcomes like achievement or attitudes to physical activity. And, given engagement involves cognitive, affective, *and* behavioural components, it is closer in the causal chain to those important outcomes than purely psychological variables like motivation or need satisfaction.

We focused on students in the first year of secondary school because student engagement with physical education typically decreases substantially during this period. For example, Barkoukis et al. (2010) found linear decreases in students' enjoyment of physical education across the first three years of secondary school. We hypothesised that—on average—students in our sample would show a similar trend. However, we also aimed to identify what teacher behaviours helped stem this decline. We hypothesised that the students' perceptions of their teacher behaviours would predict those students' change in engagement across the school year. Given many of the behaviours proposed by these theories operated by similar hypothesised causal paths, we expected many of the behaviours to be correlated with each other. Nevertheless, we hypothesised that some of these behaviours would predict whether students became more or less engaged. We had no a priori hypotheses as to whether these behaviours would be the behaviours common to those theoretical approaches, or whether predictive validity would be maximised by using the behaviours unique to each theory.

Method

Study Design

We used a longitudinal design to assess which teacher behaviours most strongly predicted change in student engagement across an Australian school year. As our primary outcome, we measured the change in student engagement between the end of their final year of primary school (Term 4 of Grade 6 in 2017) and the end of their first year of secondary school (Term 4 of Grade 7 in 2018). Students rated their perceptions of their physical education teacher's behaviour in the middle of their first year of secondary school (Term 2 or 3 of Grade 7 in 2018).

Participants and Setting

We invited all eligible government-funded secondary schools within 150 km of the Australian Catholic University, Strathfield campus to participate. Schools were invited according to their 'Index of Community Socio-economic advantage' (ICSEA) value. Schools with low socio-economic advantage (ICSEA < 1000; Australian Curriculum, Assessment and Reporting Authority, 2011) were eligible to participate. To publicise the study, we attended regional principals' meetings and conferences for physical education teachers. We also promoted the study through presentations hosted by the lead author's university and an email sent to schools' head teachers. We invited 130 schools that met inclusion criteria (median

ICSEA = 942.5, IQR = 917–972). Of those, 17 schools provided consent, with characteristics representative of the invited population (median ICSEA = 943.5, IQR = 915–964).

Within the 17 participating schools, all Grade 7 physical education teachers were eligible to participate, of which 83 provided consent to be rated by their students. All students who were enrolled in Grade 7 in 2018 were eligible except for those who did not participate in the regular curriculum. In total, 1324 students (52% female) completed at least one time point. Of students who completed the baseline questionnaire, 89% rated their teacher's behaviour and 91% completed the end-of-year data collection. Students who reported having two or more PE teachers were removed during data cleaning because we judged the questionnaire regarding the primary independent variable (i.e. teacher's behaviour) was ambiguous.

Baseline Data Collection

Once we had recruited the secondary schools, we sought to collect incoming students' data late in their final year of primary school (Term 4 of Grade 6 2017) before they entered secondary school in January 2018. To collect student data before secondary school, we either administered questionnaires at the secondary school's orientation day in December 2017, or visited the primary schools in the same geographic area (i.e. 'feeder schools'). Any students we could not assess before they entered secondary school completed assessments as soon as possible within the first term of 2018.

Teacher Behaviour Data Collection

We measured perceptions of teacher behaviour in the middle of the 2018 school year, in either Term 2 or Term 3. We aimed to collect data on a quasi-random, representative school day, while fitting in with schools' other commitments. To those ends, we asked head teachers at enrolled schools to identify weeks when data collection would be convenient. From this list, we randomly selected days upon which to collect data at the school. To reduce the risk of teachers changing their teaching approach in response to upcoming data collection, teachers were given less than one week's notice of data collection.

End-of-year Data Collection

In Term 4 of 2018, we arranged a time with the head teachers of each school during which all students could complete their questionnaires.

Variables

We collected a small number of demographic variables (gender, year of birth, perceived wealth, number of books in the home, country of birth). Descriptive statistics

and operationalisations of these variables are available on the Open Science Framework (OSF; https://osf.io/742wz/).

Pre-Post Engagement Measure

We used the Student Engagement in School measure (Lam et al., 2014), adapted to physical education, to assess three types of engagement: affective, behavioural, and cognitive. Items assessing affective engagement explored students' feelings about learning (e.g. 'I enjoy learning new things in PE class'). Items assessing behavioural engagement explored students' effort or persistence in the subject (e.g. 'If I have trouble with a skill, I try it again until I can do it'.). Students responded to affective and behavioural engagement questions on a 5-point Likert scale from 'Strongly disagree' to 'Strongly agree'. Items assessing cognitive engagement explored students' strategies to promote deep processing and retention (e.g. 'I try to understand how the things I learn in PE fit together with each other'; 5-point scale from 'Never' to 'Always'). We selected four such items assessing each component of engagement on the basis of our pilot testing. Our pilot study, conducted in late 2014 with a sample of 257 Years 7 and 8 students from Australian schools, showed that all scores from these abbreviated measures fit the hypothesised measurement models (χ^2 (243, 51) = 105.37, CFI = 0.947, TLI = 0.932, RMSEA = 0.07, SRMR = 0.050). At baseline, we worded questions to ask students about their experience during Grade 6 (i.e. '… describe your experience in Grade 6 PE practical lessons'). At follow-up, we worded questions to ask about their experience of the most recent school term (i.e. '… describe your experience in PE practical lessons over the past two months'). Our primary dependent variable was the change in these scores across the year (engagement at the end of year 7 minus engagement at the end of year 6).[2]

Teacher Behaviours

The four theories provide both overlapping and distinct recommendations for teacher behaviours. For example, all four recommend constructive, improvement-oriented feedback. On the other hand, only transformational leadership explicitly emphasises the idealised influence that comes from good role-modelling. To assess the unique contribution of each behaviour, we catalogued influential behaviours identified by these theories. Then, we distilled this catalogue into a set of distinct behaviours. Finally, we sought to assess the extent to which students perceived that their teachers displayed these behaviours in their lessons. These steps are outlined below.

To create a comprehensive list of behaviours, we used a binning and winnowing protocol (DeWalt et al., 2007; Mâsse et al., 2016). This process has been widely used in health research. For example, Mâsse and colleagues (2016)

---

[2] While there is ongoing discussion as to whether difference scores or residualised change scores are more reliable and valid (Williams & Zimmerman, 1996), we judged that difference scores were appropriate in this context because modelling studies have shown 'the difference score model is a better choice in nonrandomized studies' (Castro-Schilo & Grimm, 2018, p. 47).

established comprehensive sets of physical activity parenting practices, while DeWalt et al. (2007) identified patient-reported outcomes from chronic diseases. We used four key steps: identify the pool of candidate behaviours; group similar behaviours into categories (i.e. binning); winnow behaviours from categories into an exemplar; and refine examples through member checking.

To ensure coverage of motivational behaviours, we first created a pool of teacher behaviours assessed in the literature. We searched for questionnaire items that assessed teacher behaviours, either from the perspective of the student, from an observer, or from the teacher themselves. We identified relevant studies by reviewing the reference lists of two recent and relevant systematic reviews on teachers' motivational behaviours in physical education (Smith et al., 2016) and other school subjects (i.e. Science, Technology, Engineering and Mathematics; Rosenzweig & Wigfield, 2016). We also reviewed the reference lists of the papers included in those reviews. We also searched for more recent studies using Scopus and Google Scholar until saturation: until new measures appeared to add no new items to those in the pool. From 111 papers, we sourced 49 unique scales, containing 1151 candidate items.

The next step was designed to cluster these candidate items into categories ('binning' process; DeWalt et al., 2007; Mâsse et al., 2016). Five authors created a list of 48 categories based on theory (e.g. 'Offering input or choice'). Eight authors then each took a random sample of 260 candidate items (e.g. 'provides meaningful choices'; Smith et al., 2016) and placed them into categories. Each item was placed into 'bins' of similar items by two authors, independently and in duplicate. When an item did not fit any category, authors inductively created a new category for the behaviour. This process led to an expanded list of 61 categories that covered all teacher behaviours, with a list of candidate items in each category.

We sought to identify a single exemplar item in each category by winnowing the list of candidate behaviours (DeWalt et al., 2007; Mâsse et al., 2016). This process involved deleting redundant, narrow, or confusing items until there was one item for each category. We also drafted examples for each item where doing so provided clarity (e.g. for providing a transparent structure for students: 'In today's class, we're working on throwing skills in three ways…'). This process was initiated by one author and reviewed by another. Exemplar behaviours were then member-checked by two groups: the eight authors who categorised items and five teachers from local secondary schools. These two groups identified behaviours that were unclear or redundant, and were asked to identify any other motivational behaviours that were not covered by the categories. At the end of this process, we identified 69 unique motivational teacher behaviours, outlined in Supplementary Table 1. These behaviours are largely similar to the final list derived by three-round Delphi process in collaboration with 34 experts in self-determination theory (Ahmadi et al., 2023).

We then classified each motivational teacher behaviour under one or more of the motivational theories. Four authors collaboratively classified each behaviour as characteristics of each theory. For example, behaviours were classified as consistent with achievement goal theory if the behaviour was consistent with the theory's TARGET framework (Jaitner et al., 2019). We did the same

classifications for transformational leadership theory and mindset theory. The results of this classification are also available in Supplementary Table 1.

This list of behaviours appeared to cover most concepts within all four theories, except idealised influence of transformational leadership. Idealised influence is less conducive to assessment through observable behaviours, and is often measured as a student perception of a leader characteristic (Beauchamp et al., 2010). To that end, we also asked students to rate teachers on two such characteristics to ensure full coverage of the concept ('My teacher acts as a person that I look up to' and 'My teacher behaves as someone I can trust'; Beauchamp et al., 2010). As a result, our final list included 71 unique behaviours, 68 related to a psychological need in self-determination theory, 47 of which were characteristic of a leadership style in transformational leadership theory, 29 of which were characteristic of either a mastery or performance climate, and 10 were likely to influence students' growth or fixed mindset. Students rated teachers on all 71 motivational behaviours. For all items, students responded on a 7-point Likert scale with descriptors at each point from *Strongly Disagree* to *Strongly Agree*. Our primary independent variable was, therefore, student perceptions of their teacher's behaviour—rather than objectively measured behaviour from lesson observations—because we felt it was more sensitive to differences in the way teachers treated each student (or in how students experienced their teacher's behaviour).

Bias

We anticipated social desirability and common method variance as the most likely sources of bias. Students may have felt pressure to describe their teacher more favourably if teachers were administering questionnaires. For this reason, students completed their questionnaires under the supervision of the research staff. They also completed the descriptions of teacher behaviours at different times to the measures of engagement, with most students having at least 4 months between each time-point (see Fig. 1).

Data Analysis

Missing Data and Preprocessing

We imputed missing data on predictors with K-Nearest Neighbours (Tutz & Ramzan, 2015) using caret (Kuhn, 2008) in R (R Core Team, 2020). We generated a model for imputing missing data using the training data, then applied this model to the test data to avoid data leakage (Jaeger et al., 2020). Data leakage is where patterns in the testing data are used to inform the training data. When that happens, the models created on the training data may learn patterns about the test data. This violates assumptions in the test data—that the data is unseen in training. Predicting the test data is the focus of predictive models, rather than
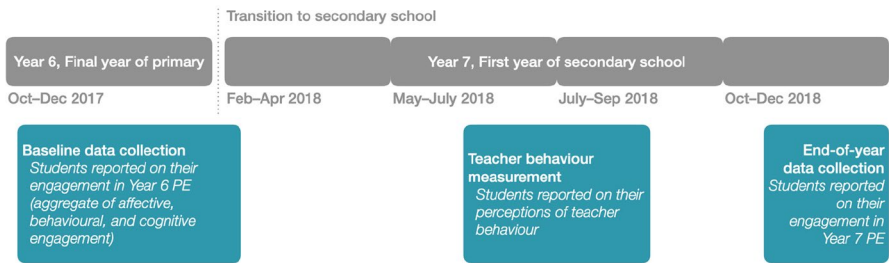
Fig. 1 Study design, data collection points, and variables

trying to make inferences about the latent parameters that explain an association. As a result, it is important to be able to reproduce the 'imputation generator' for imputing missing values on the test data (Kuhn & Johnson, 2019). When using standard methods for inferential statistics, like multiple imputation, it is not possible to extract the model for imputing missing data on the test set. Because the primary objective of predictive models is predictive accuracy on the test set, it is also important to approximate the missing data as closely as possible, rather than trying to preserve uncertainty in the predictions (e.g. via multiple imputation; Kuhn & Johnson, 2019). K-Nearest Neighbours is an efficient and effective model for imputing modest amounts of missing data (Tutz & Ramzan, 2015), as is the case here (11% missing). This imputation method identifies the K (usually 5–10) most similar cases to one with the missing data and imputes the missing value using the mean of those cases (Kuhn & Johnson, 2019). The model can be trained on the training data and used to impute missing data in the test data without violating assumptions of the test set.

In addition to imputing missing values, we also scaled and centred each predictor, as recommended when using our predictive models (i.e. elastic net; Hastie et al., 2009; James et al., 2013). Scaling and centering sets the means to 0 and the variance to 1 for all predictors. Doing so is important for an elastic net because the model is penalised based on the size of each coefficient. Without scaling the predictors, some predictors would be penalised because the predictor had a high variance, even if it was an excellent predictor for the outcome. We did not control for clustering, for three reasons. First, the school-level ICC was low (ICC = 0.04). Second, doing so would reduce the generalisability of the predictive models so they could not be used with schools beyond the training set; if school was entered as a variable, then researchers could only use the model in the schools for which we built the models. Nevertheless, an exploratory analysis that did centre all predictors within clusters (Usami et al., 2019) did not lead to any meaningful changes in beta values or predictive validity (see Exploratory Analyses at https:// osf.io/742wz/). We did not account for clustering within teachers because differences between teachers is an important source of variance in our independent variable.

Descriptive Methods

The binning and winnowing procedure was designed to ensure that no two items would measure the same behaviour (i.e. behaviours would be mutually exclusive). Those items would have ideally ended up in the same 'bin' and would, therefore, have been collapsed into a single item. To assess whether this was successful, we calculated inter-item correlations as a measure of redundancy (Cohen et al., 2012). We considered items with correlations above 0.7 (or below -0.7) redundant, but no items met this threshold ($r_{max} = 0.68$). Declines in student engagement across the year were quantified as standardised mean differences using Cohen's *d* via the *effsize* package (Torchiano, 2020) in R (R Core Team, 2020).

Predictive Methods

Typically, researchers using explanatory models do not use 71 different predictors in a single statistical model. This is usually for three reasons (Hastie et al., 2009). First, as described in the introduction, models with dozens of predictors are likely to 'overfit' the training data. The models will learn spurious correlations that do not translate well into new data. Second, these models are much harder to interpret. Given a beta value represents the partial correlation between the predictor and an outcome, the beta for a regression between y_1 and x is the influence of y_1 on x while controlling for y_2, y_3, all the way to y_{71}. As will be demonstrated later, this can lead to results that are difficult to interpret. Finally, most linear models have an assumption that there is no multicollinearity between predictors. This is largely because multicollinearity exacerbates the problems above. For example, if two predictors were very highly correlated, then the beta values for each represents the predictive influence from the small amount of unique variance not shared by the two predictors. This makes the predictors unreliable and relatively uninterpretable.

In explanatory models, these problems are usually managed by using dimension reduction techniques that aggregate multiple items into a latent construct (e.g. factor analysis, principal components analysis; Hastie et al., 2009; James et al., 2013). These techniques are useful in explanatory research because they are generally trying to explain the role of the latent construct in predicting the outcome (Shmueli, 2010). The minor variations in people's responses to each item is somewhat dismissed as measurement error. The problem with these techniques is that they obfuscate the influence of any particular item, with some items loading on multiple dimensions, or loading to various degrees (Shmueli, 2010).

With predictive models, we are not constrained by using a small number of predictors representing latent factors. Because the predictive culture is attuned to the risks of overfitting, and is less concerned with explanation than predictive accuracy, researchers will often use more complex methods and explicitly use strategies that avoid overfitting. Many of these methods allow researchers to model thousands—or even millions—of predictors while also identifying which of the specific predictors are most influential. This means that rather than aggregating teacher behaviours into latent constructs, we could treat each behaviour as a separate independent variable. So we could determine the predictive contribution of unique behaviour, we used an

elastic net to manage overfitting and multicollinearity, while also allowing for easy interpretation of the contribution of individual behaviours (Zou & Hastie, 2005).

An elastic net is an ordinary least squares regression model with some penalisation terms that helps manage overfitting and collinearity. One form of penalisation involves shrinking the *size* of regression coefficients, particularly those that are collinear (ridge regression; Kirpich et al., 2018). Where normal 'least squares' regression aims to minimise the squared residuals (hence 'least squares'), ridge regression aims to minimise the squared residuals plus the squared coefficients, which penalises large coefficients. This means large coefficients that do not reduce error in prediction are shrunk, such as when two predictors are multicollinear. For example, if *clear instructions* and *unclear instructions* were intercorrelated with the outcome, both could be used to predict the outcome but their coefficients would be shrunk. Another form of penalisation involves mostly shrinking the *number* of coefficients used to predict the outcome (lasso regression; Kirpich et al., 2018). In lasso regression, the penalty term is based on the absolute value of the beta values, meaning larger coefficients are not necessarily penalised, but inefficient predictors are just shrunk to zero. In the above example where two predictors were collinear, then *clear instructions* might be included in the model and the coefficient for *unclear instructions* would be reduced to zero. Elastic net regression allows for both of these penalisation terms (Zou & Hastie, 2005): it can shrink and select coefficients, depending on which leads to better prediction in the data.

To determine whether it is better to shrink the size or number of coefficients, we used a data-driven approach through cross-validation, recommended by machine learning researchers (Hastie et al., 2009; James et al., 2013; Zou & Hastie, 2005). We used a random number generator to split our data, row-wise, into a training set (70%) and a testing set (30%). That is, of the 1324 participants in our sample, we randomly selected 30% of these participants ($n_{test} = 398$) to be held out from training. Because the test set was randomly sampled using a number generator, any differences between test and training sets were due to sampling variance. Differences between groups on continuous variables were negligible ($d_{mean} = 0.05$), and there were no statistically significant differences on categorical variables (all $p > 0.05$). In this test set, we only assessed model fit on these participants once we had finalised our parameters (e.g. beta values) and hyperparameters (e.g. amount of penalisation). On the training set ($n_{training} = 926$), we used fivefold cross-validation to tune both hyperparameters for *glmnet* (Friedman et al., 2022) using *caret* (Kuhn et al., 2022) in *R* (R Core Team, 2020); that is, we repeatedly used the training data to generate models, varying the parameters for both the type of regularisation (shrinking parameters vs. mostly eliminating them; the α parameter) and the degree of regularisation (strong vs. permissive; the λ parameter). We used the parameters that minimised the error in predicting engagement during cross-validation.

Once parameter tuning was complete, we tested the model on the unseen data—the test set—to see how well the model would likely generalise to future samples of physical education teachers. We used prediction R^2 to measure the overall model fit, and beta-values for each individual predictor. Prediction R^2 is slightly different from the R^2 generally used in explanatory models (Scheinost et al., 2019). R^2 from explanatory models reflects the *correlation* between the

predicted and observed values. Predicted and observed values can be highly correlated while still being very inaccurate, such as when all predictions are systematically inflated. Therefore, prediction R^2 is preferred because it assesses the *error* between predicted and observed values, as a proportion of the error from merely predicting the mean. It can therefore be negative if the model is worse than merely predicting the mean.

Statistical significance is not meaningful in the context of an elastic net. This is because the process of penalising coefficients means ineffective or redundant predictors are eliminated, and evaluating the data on a test-set is a stronger assessment of generalisability than any inferential statistics (Hastie et al., 2009; James et al., 2013; Zou & Hastie, 2005). For comparison, we also present the results of a traditional linear model using the same data. Some readers may want models adjusted for covariates (e.g. gender, wealth), such as when designing interventions, so that models explain the residual variance once adjusting these variables. While not the primary goal of our manuscript, we present adjusted models on the OSF (https://osf.io/742wz/). The covariates explained some data in the training set and did not increase test-set performance, and most predictive behaviours were unchanged. For this reason, only our primary, unadjusted models are presented in the manuscript. The data and code required to reproduce all analyses are available at https://osf.io/742wz/.

Results

Student Engagement in Physical Education Declined Over the First Year of Secondary School

As expected, total engagement declined across the school year by 0.25 standard deviations (95% CI = [0.17, 0.33], see Fig. 2). The same pattern was true for affective engagement ($d = 0.27$, 95% CI = [0.19, 0.35]), behavioural engagement ($d = 0.23$, 95% CI = [0.15, 0.31]), and cognitive engagement ($d = 0.14$, 95% CI = [0.06, 0.22]). As seen in Fig. 2, while the trend is toward a decline, some students report a significant increase in engagement over the school year.

Mid-year Reports of Teacher Behaviour Correlate with Changes in Engagement

As shown in Supplementary Fig. 1, a large number of teacher behaviours showed significant raw correlations with change in engagement. For example, when students perceived their teachers as good role models, they saw moderate increases in engagement over the school year ($r = 0.22$, $p < 0.001$). At the other extreme, when teachers were seen to provide unfair rewards, students reported significant decreases in engagement ($r = -0.12$, $p < 0.001$).
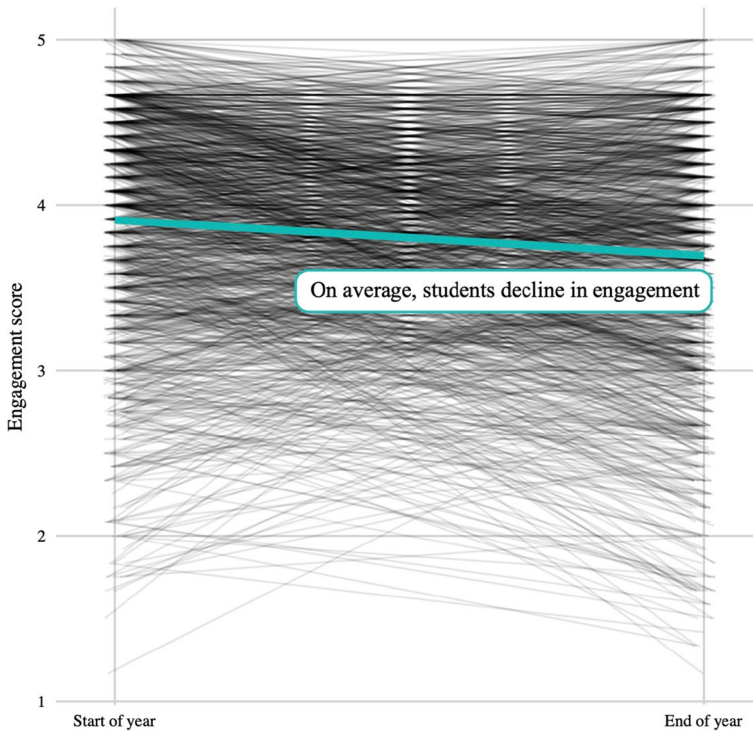
Fig. 2 Change in physical education engagement across the first year of secondary school

Teacher Behaviours Were Correlated but not Redundant

As seen in Supplementary Table 1, the student-reported teacher behaviours were correlated with each other, but were not redundant. The average correlation between behaviours was 0.21. The largest positive correlation was between ratings of teacher enthusiasm and warmth ($r = 0.68$, $p < 0.001$). The largest negative correlation was between the amount of abusive language students heard from teachers and the amount of attention they saw teachers giving ($r = -0.3$, $p < 0.001$). These data suggest that we achieved our goal of creating a mutually exclusive set of behaviours. However, with predictors correlated as high as 0.68, a standard multiple regression or linear mixed models would approach multicollinearity (Dormann et al., 2013), so variable-selection models were appropriate.

Lasso Regression Identified a Series of Predictive Behaviours on Unseen Data

In cross-validation, the glmnet algorithm consistently identified that lasso regression was a better fit for the data than ridge regression. That is, the variance explained by the model was higher when managing multicollinearity by eliminating predictors ($R^2 = 5.09\%$) rather than shrinking them ($R^2 = 3.45\%$) or any combination of the two.

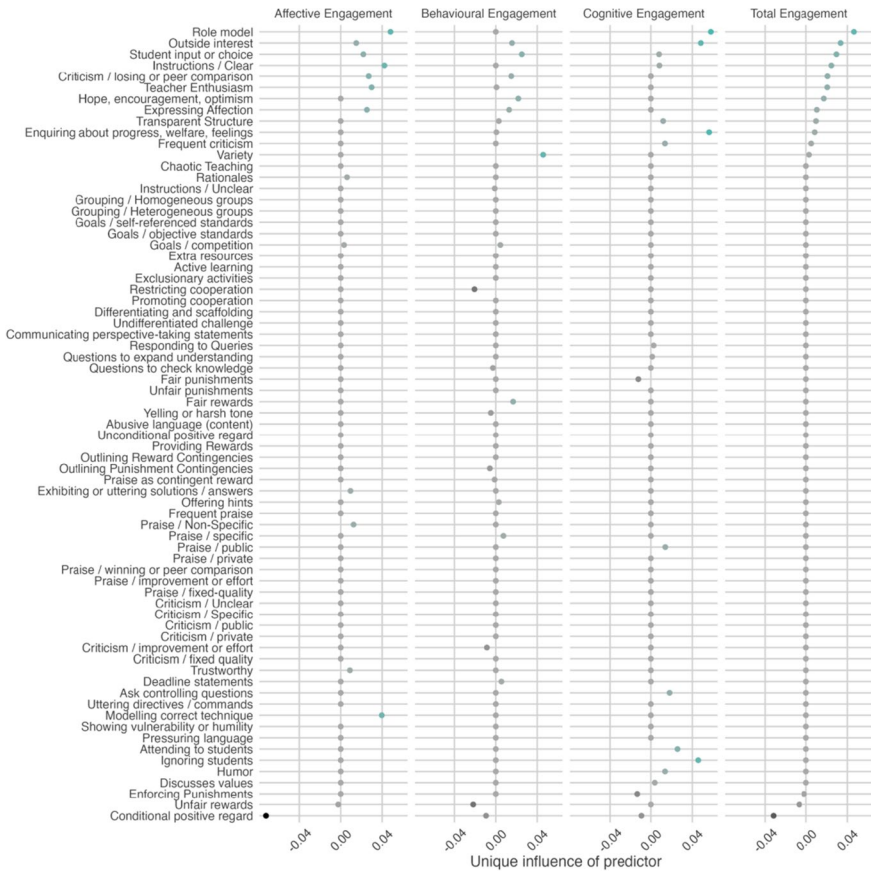*Predictive Estimates Teacher Behaviours Predicting Engagement*

Fig. 3 Predictive estimates teacher behaviours predicting engagement

Specifically, in cross-validation, the best fitting elastic net models always used $\alpha = 1$, corresponding to lasso regression (Hastie et al., 2009; James et al., 2013; Zou & Hastie, 2005).

On the testing data, the models explained 4.36% of the variance in change in total engagement. This was only slightly smaller than the variance explained in the training data ($R^2 = 5.09\%$). This finding suggests the penalisation successfully prevented the models from over-fitting to that data, and are likely to generalise well. The models were somewhat less accurate when predicting the subcomponents of engagement: affective engagement ($R^2 = 2.83\%$), cognitive engagement ($R^2 = 3.49\%$), and behavioural engagement ($R^2 = 2.92\%$). Figure 3 shows the most predictive behaviours for all outcomes. Controlling for the influence of all other teacher behaviours, students became more engaged when they felt their teachers were good role models ($\beta = 0.046$), when they took interest in students' lives outside of class ($\beta = 0.033$), and when they allowed students to make choices ($\beta = 0.029$). Students became less

engaged then they felt their teacher's regard for them was conditional upon good behaviour ($\beta = -0.032$). As outlined in Supplementary Table 1, the beta values for the majority of behaviours was zero, meaning those behaviours were not necessary to optimise prediction using an elastic net. As outlined below, using those predictors led to poorer prediction on unseen data.

Comparison with Ordinary Least Squares Regression

For comparison, we used the training data to produce a simple linear regression. Without bootstrapping or cross-validation, the model fit the training data very well ($R^2 = 13.88\%$; $R^2_{\text{adjusted}} = 6.72\%$). As expected, however, the model fit the test set far less accurately ($R^2 = -0.86\%$). A negative predictive R^2 means that the simple linear regression models were less accurate than merely predicting the mean of the outcome, which is a symptom of over-fitting. This overfitting was not present in the regularised model (lasso, described above) built using cross-validation (where $R^2 = 4.36\%$). In addition, many beta-values in the linear model are surprising (Supplementary Fig. 2). For example, the coefficients were almost identical for both conditional positive regard ($\beta = -0.059$) and unconditional positive regard ($\beta = -0.046$), despite being effectively opposite behaviours. These coefficients can occur like this because, in regression, betas represent partial coefficients. These coefficients look at the unique influence of this behaviour while controlling for all other behaviours: 'for teachers whose students believe them to be identical in every other way, the teacher that does [this behaviour] more raises engagement'. Given many variables were highly correlated with each other (Supplementary Table 1), the unique variance explained by each behaviour can be small. This is a good example where 'feature reduction' techniques from predictive models—where the number of predictors is reduced empirically, like in the lasso model described above—can lead to better prediction while maintaining or even improving interpretability.

Discussion

In this paper, we focused less on the traditional inferential statistics used to *explain* educational psychology phenomena, and instead drew from the cultures and best-practices for *predicting* those phenomena (Breiman, 2001; Shmueli, 2010; Yarkoni & Westfall, 2017). We used a longitudinal design to identify the teacher behaviours that most predicted changes in student engagement during physical education. We also aimed to assess how well the behaviours of four influential educational psychology theories (self-determination theory, transformational leadership theory, achievement goal theory, and growth mindset theory) cumulatively predicted student engagement in physical education in low SES schools. Consistent with previous research, we found that student engagement declined across the first year of secondary school. We found that a mutually exclusive and collectively exhaustive set of behaviours drawn from the four theories predicted ~5% of the variance in students' change in engagement. To do so, the best predictive models tended to draw from a

relatively small number of behaviours, including both helpful behaviours (particularly quality role modelling and taking interest in students) and unhelpful behaviours (conditional positive regard). As shown in Supplementary Table 1, impactful behaviours were aligned with both self-determination theory and transformational leadership theory, but seldom explicitly aligned with the core tenets of achievement goal theory or mindset theory.

The percentage of variance in change in engagement we explained is only moderate compared with typical benchmarks (Funder & Ozer, 2019). A correlation of 0.2 ($R^2 = 4\%$) 'indicates an effect of medium size that is of some explanatory and practical use even in the short run' (Funder & Ozer, 2019, p. 166). This may appear modest given the number of teacher behaviours we examined; however, it is important to consider these effects in the context of three important considerations. First, the turbulent context of our study (i.e., transition to high school) and the longitudinal research design mean we may need to be more modest in our expectations for how much variability in student experience we can explain by just the student's perceptions of their teachers. Second, we saw that typical models from explanatory statistics, like multiple regression, can show very different predictive accuracy (R^2) on the data used to build the model and new data used to test it. The R^2 for our linear model dropped from 13.88 to -0.86% when testing the model on new data, meaning they were worse than merely predicting the average change in student engagement. This drop is likely to be more dramatic in our models than would be typical in educational psychology because of the number of predictors we included. Nevertheless, it highlights the risks that arise from testing models on the data used to train them: researchers may develop an illusion of predictive accuracy. Some typical, explanatory models may be overstating their predictive accuracy by never testing their predictions on new data. Shrinking or penalising the size of coefficients, as we did in our lasso regression, can make predictors look less powerful, but they actually lead to better predictions on new data. Finally, our penalised model showed that we could predict change in engagement reasonably well while only using a handful of discrete teacher behaviours (~10). Each of these behaviours may only modestly predict engagement: for example, the beta (β) values showed that a standard deviation increase in showing interest in students outside school led to a 0.035 increase in the student's predicted engagement. However, showing interest in students outside of school is only one of hundreds of behaviours that teachers may do to increase engagement, and these effects are controlling for the influence of those other behaviours. As Funder and Ozer (2019) note, even small effects like these are 'potentially consequential in the not-very-long run' (p. 166).

Interpretations of Predictive Results

As described in the introduction, errors in prediction help researchers identify gaps in theoretical models. Educational psychology researchers may want to explore reasons for the predictive accuracy of the models found here. As described above, by some standards, our models predict change in engagement moderately well. However, our model attempted to use the combined recommendations from four major

theories in educational psychology. As a result, others may view the predictive accuracy as disappointing, but as noted above, few studies used comparable methods for testing their predictive accuracy (i.e. using a hold-out test data set). Without researchers more frequently using predictive methods like those presented here, it is not possible to robustly compare the quality of theoretical predictions. If the field more frequently assesses the predictive validity of their models, we might more quickly find what we might be missing from our models, so the field can devise better explanations and better integrate different theories.

It is important to robustly assess how well psychological models predict outcomes we care about (Yarkoni & Westfall, 2017), but weak predictive accuracy does not necessarily rule out a causal relationship. If our model did not select a variable, it does not mean it has no connection with the outcome. Our predictive models can help researchers predict whether a student is likely to become more or less engaged, based on their perception of their teacher on a few key behaviours, but they do not allow us to confirm that those behaviours are the strongest causal factors influencing engagement. Role modelling may be a strong predictor of engagement because it is so highly correlated with other things students find motivating, like teachers talking about their values or showing students warmth. It is possible those behaviours are the causal drivers, or that some third variable (e.g. teacher's age) influences the student's perceptions of all three. As with many observational designs, predictive behaviours from these models do not necessarily mean those behaviours causally influence engagement.

Similarly, bad predictors are not necessarily unimportant. As mentioned earlier, both mindset theory and achievement goal theory have systematic reviews demonstrating how constructs from these frameworks influence students (Biddle et al., 2003; Jaitner et al., 2019; Ntoumanis & Biddle, 1999; Sisk et al., 2018). One explanation for why those behaviours were not consistently selected in our models is that the effects may be relatively smaller than those of other teacher behaviours (Sisk et al., 2018). Or those behaviours might have longer structural causal models (Pearl, 2010), whereby teacher behaviour less directly influences engagement. The more 'dominoes that need to fall', the smaller the chance that influencing a distal cause will lead to the hypothesised outcome. As an analogy, socioeconomic status is often less predictive of university enrolment than academic achievement (Parker et al., 2018), but that does not mean economic status is an unimportant explanatory variable. It is just more likely that socioeconomic status predicts university enrollment because of its effect on achievement. In our study, teachers setting mastery goals may increase engagement, but only if doing so *changes* student goals from performance to mastery, and only if making such a change helps students feel competent in a situation where they would have otherwise felt failure. If students already held performance goals, or if they felt a sense of failure anyway (despite the mastery goal), then the teacher's efforts to influence the goal climate may have been a weak influence on engagement.

Alternatively, behaviours may have also been weak influences because of low variability between teachers. For example, most researchers would expect 'abusive language' to have a strong negative effect on engagement. However, teachers calling students hurtful names may be rare, so students may not reliably describe how

frequently it happens across the year. It may have been a weak predictor because, thankfully, teachers know this is a damaging practice and genuinely avoided it. As a result, this may be a poor predictor and a weak target for an intervention in this context. In contrast, teachers have a more complicated relationship with behaviours like 'conditional positive regard'. On one hand, removing warmth and attention is often recommended to teachers as a behaviour management strategy on the basis of operant conditioning (Gable et al., 2009). On the other hand, self-determination theory proposes that conditional regard can thwart all three psychological needs, making students feel less competent and pressured to conform, while damaging the teacher-student relationship (Kanat-Maymon et al., 2021). Because of these competing arguments, teachers may be more likely to vary on a behaviour like conditional positive regard, both between teachers and in the way they treat individual students. Due to that variability, it may be a strong predictor of engagement and a particularly useful target for intervention.

In summary, some behaviours with likely causal effects on engagement (e.g., abusive language, achievement goals) may not show up in predictive models. These findings highlight a trade-off between predictive and explanatory methods—while better causal understanding often improves prediction, better prediction does not always improve causal understanding. Still, predictive models can help researchers identify high-leverage variables that most strongly predict changes in engagement. These predictive models can be used now to assign resources or identify teachers in need of further training while we search for explanatory models with better predictive performance.

These Behaviours May Be Good Focal Points for Interventions to Improve Engagement

As described above, behaviours that predict changes in engagement may not necessarily be the strongest causal influences. Our observational design and predictive models cannot make strong causal claims, almost by design. To determine whether these behaviours causally influence student engagement, other designs are necessary (e.g., a randomised trial where some teachers are encouraged to adopt the behaviours that predict higher engagement).

Nevertheless, it is possible that the stronger predictors of engagement found in our data may represent shorter causal models, higher variability between teachers, or more powerful effects. The behaviours selected by our model may therefore be useful components of interventions for increasing PE engagement among low-SES students, or may be useful flags for teachers who may benefit from additional training (e.g. those who students feel are poor role models). Indeed each of the most predictive behaviours have clear theoretical underpinnings and robust data for their importance. For example, in contrast with small effects from growth mindsets (Sisk et al., 2018), meta-analyses have consistently shown that teacher-student relationships have strong direct effects on engagement (Roorda et al., 2011, 2017). As a result, it may not be surprising that engagement in physical education would be influenced by behaviours that directly influence those relationships (e.g. positive role modelling,

taking interest in students, conditional positive regard). Providing choices is a central recommendation of autonomy supportive interventions, which have been shown to reliably increase engagement (Reeve & Cheon, 2021). Similarly, the benefits of clear instructions can be understood through cognitive load theory (Sweller et al., 2019), where poor instructions lead to cognitive overload. Dozens of meta-analyses have shown that interventions to improve the clarity of instruction improve student learning (Noetel et al., 2021). So, many of the features selected by our predictive model have meta-analyses supporting their use and plausible causal chains explaining them. However, the same could be said about the dozens of other behaviours hypothesised to increase engagement. Predictive models may help researchers choose from those behaviours, then testing those behaviours in experimental designs would then provide stronger causal evidence of their effects. This interaction shows how predictive and explanatory statistics can complement each other.

Data-Driven Decisions May Be Well Informed by Predictive Models

The influential behaviours outlined above could each be *explained* by compelling psychological theories. But, by simultaneously assessing the *predictions* of multiple theories, we could assess their relative importance. Because the majority of educational psychology literature has focused on explanation, it has not been able to disentangle which specific behaviours are most predictive. Randomised controlled experiments, and meta-analyses of these experiments, have almost always 'bundled' a set of theoretically aligned behaviours (Lazowski & Hulleman, 2016). Instead, by breaking apart the theories into the respective behaviours, and simultaneously assessing the influence of each behaviour, we were able to identify the behaviours that might be the highest leverage.

This leverage is important because some theories predict *so many* teacher behaviours would be influential. Our binning and winnowing procedure identified 71 behaviours, almost all of which could be conceptualised within self-determination theory. Being a comprehensive meta-theory, SDT does provide compelling *explanations* for why these behaviours may be important, but it does not provide clear *predictions* for which behaviours would be most important. For example, one of the most predictive behaviours in our data was the teachers' role modelling. While SDT may provide explanations for why role modelling is important, the behaviour itself receives more emphasis in transformational leadership theory (Bass, 1985) and social identity theory (Rees et al., 2015). No one theory emphasises all the most important constructs, because as identified in other articles of this special issues, there is a 'goldilocks zone' of theoretical breadth (Dinsmore et al., n.d.): too narrow and the model is not predictive; too broad and it is imprecise. Cross-theoretical approaches to model creation and intervention design may help researchers balance these competing forces of specificity and sensitivity. We want our models to include all the most important behaviours that influence the outcomes we care about (sensitivity). We also do not want to waste time and resources teaching skills that only weakly influence outcomes (specificity). Studies like ours demonstrate methods

through which researchers can make data-driven decisions about how to navigate this trade-off.

Unseen Data Predictive Accuracy Shows Opportunities for New Methods and Models

We used a robust and reproducible method for creating a mutually exclusive and collectively exhaustive list of teacher behaviours drawn from four theories of motivation. As hypothesised, many of these teacher behaviours were correlated with changes in engagement, but due to high correlations with each other, typical regression or dimension reduction approaches were not appropriate. Instead, we used an interpretable machine learning approach (elastic net regression) to assess the combined predictive validity of these behaviours on unseen data. Given the longitudinal design with spaced measurement, these models explained a practically significant percentage of the variance in change in engagement ($R^2 = 4.84\%$). This is a substantial percentage of variance given the plethora other important factors that would influence student engagement across a year, such as learning materials and peer influences (Hattie, 2008). This may be particularly true in the first year of high school where students are entering an entirely new environment. In other words, it is reasonable to have modest expectations of the influence teachers have on their students. However, many educational psychologists might be disappointed with this level of predictive accuracy from the collective suggestions of four established theories.

Using norms from psychology and education (e.g. Funder & Ozer, 2019), our pooled predictive accuracy would be considered moderate at best; however, those norms were established using methods for testing model fit on the 'seen' data alone. So, it is possible that our R^2 is only moderate because of our conservative study design, but it is more likely that most other studies are overstating their predictive accuracy by calculating R^2 on seen data (Yarkoni & Westfall, 2017). R^2 is almost always higher on seen data than unseen data (James et al., 2013), but the size of that gap is important to quantify if we want to improve our models. As we saw from our ordinary linear regression, models can appear to fit well on the data used to train them, but rather useless on new data. We are sure this is an extreme example, and are not suggesting the same would be true of previous educational psychology research, but it serves as a useful warning example. As far as we know, our study is one of a small handful in educational psychology that assessed the predictive accuracy of their models on unseen data (Parker et al., 2018), so the amount of 'overfitting' in educational psychology is hard to assess. We hope more educational psychology researchers test their models prospectively on new data, or at least test them retrospectively on some quarantined test data as we did here.

We hope this paper functions as a call-to-arms in educational psychology, identifying a significant gap in the field and a pathway for improving it: given some complex models can only explain a modest amount of change in an important educational outcome, how can we update our models to improve our predictive accuracy? And when we do, can we become more comfortable with lower R^2 values by embracing the best parts of the culture of prediction (e.g. testing models on unseen

data)? This may feel unnerving for researchers, but if we want to assess the generalisability of our theories, it is important to quantify how well models work on data not used to build the model. Only by trying to increase our predictive validity, *as well as* our explanatory power, can we, as educational psychologists, meet the goals we set for ourselves as a field. We suggest that a combination of both approaches will be superior and complementary, rather than an overwhelming reliance on explanatory models.

Strengths, Limitations, and Future Directions

Prediction and explanation are always in tension (Shmueli, 2010). By focusing on prediction, we chose to spend less time explaining. Studies of mediation analyses have been able to show that, for example, need satisfaction and self-determined motivation might explain why interventions grounded in transformational leadership theory work well (Castillo et al., 2020). Our study was not designed to test these hypotheses, which is unorthodox given the explanatory culture of fields like educational psychology (Breiman, 2001). While each approach, explanation and prediction, has strengths, researchers must make compromises between the two. More complex, flexible models tend to have better predictive accuracy but also tend to be harder to interpret. Researchers continue to advance explainable artificial intelligence methods, which aim to get the best of both worlds: the predictive accuracy of complex models with the ability for humans to explain how models work (Phillips et al., 2021). We used an elastic net because it is a relatively interpretable machine learning method of managing multicollinearity, improving predictive accuracy while maintaining interpretability (Molnar, 2021). Other machine learning methods may allow for more complex models with a larger number of non-linear or interaction terms, but they are less interpretable than an elastic net, which provides familiar β estimates. Still, there are limitations to these methods. By shrinking highly correlated predictors, the models can reduce to zero predictors that may be important but that are just better explained by nearby behaviours that are correlated. For example, expressing warmth and affection was correlated with being a good role model (Supplementary Table 1), but affection was frequently (but not always) shrunk to zero (Fig. 3). When it was selected, it demonstrated relatively strong effects, but this would likely account for the few times where 'role-model' was not selected. In predictive models, it is not possible to account for this variability between models without introducing biases (Hastie et al., 2009; James et al., 2013; Yarkoni & Westfall, 2017). These are necessary compromises that come with the trade-off between predictive accuracy and interpretability.

Our study was a real-world assessment of a sample across 17 lower socioeconomic advantage schools. The longitudinal design allows for stronger causal inference and management of common method variance than cross-sectional designs. However, it was still observational and relied on student reports of how they perceived their teachers, so may be subject to the characteristic risks of bias and noise. For example, a strength of our design was assessing most students before they had encountered their Year 7 physical education teacher. However, doing so also means we measured students before and after one of the most significant transitions in a students' journey. As a result, there may be substantial amounts of noise introduced into students'

changes in engagement that may be explained by plethora other factors (e.g. new friends, larger school, new structure, all new teachers, moving from being the oldest in the school to the youngest). Also, any students we could not assess before they entered secondary school completed assessments as soon as possible within the first term of 2018. This means some of the students' pre-test ratings of Year 6 engagement may have been influenced by their early experiences of their Year 7 teacher. We also could not collect data from all 17 schools at the same time, so the interval between our measurement of teaching practice and the outcome differed between schools. We judged that this artefact would create random variation that would be unlikely to systematically influence the predictive models—in fact, we assume this random variation makes the models more generalisable—but it plausibly reduced our predictive accuracy. Similarly, the specific sport or activity they were studying might be a random variable that would have induced noise at each timepoint. For example, some activities might be less conducive to cooperation or engagement than others (e.g. gymnastics vs. volleyball). Our statistical methods were designed to assess the generalisability of models to other samples, but given the data were all collected from low-SES Year 7 physical education classrooms, the predictive model may not generalise to other subjects (e.g. science, literacy) or less disruptive developmental periods (i.e. compared with the transition to secondary school). Nevertheless, we think the *methods* we use here—particularly the predictive models being tested on unseen data, not used to train the model—could be used more widely throughout educational psychology. Doing so would help us find gaps in our theories and inform useful ways to synthesise those theories.

Finally, the taxonomy of teacher behaviours we created led to a mutually exclusive set of behaviours, but we cannot guarantee they were collectively exhaustive. Consensus building exercises with larger panels, like Delphi studies, are the gold standard for creating these taxonomies, but doing so was beyond the scope of this paper. The behaviours identified in this paper were subsequently refined through a three-round Delphi study with a panel of experts in self-determination theory (Ahmadi et al., 2023). Although most of the final behaviours matched those in our list, those experts provided suggestions that were not available in time for use in this study. Our specific items and examples, therefore, may not have been the optimal way of asking students about these behaviours. Similarly, we chose four theories from the many good options (Lazowski & Hulleman, 2016; Skinner & Pitzer, 2012). For exploring student engagement, any choice of four theories will omit theories that make useful, unique contributions. While we attempted to draw the key hypothesised behaviours from each of the four theories, and attempted to choose diverse theories of interest to researchers and educators, we support ongoing efforts to compare, contrast, and develop theories that better predict and explain student engagement.

Conclusions

Many educational psychological theories aim to predict the influence that teachers will have on their students. In this paper, we used methods from the predictive culture of statistics to see which teacher behaviours best predicted engagement in

physical education. We simultaneously assessed the predictions of four motivational theories, and also assessed the cumulative ability of those theories to predict change in engagement as students transitioned into secondary school. While the combined behavioural predictions from those four theories explained a modest percentage of variance on unseen data, a small number of behaviours consistently predicted student engagement. Physical education teachers can engage students by being good role models, discussing class values, taking interest in students' lives, differentiating lessons, and by avoiding unfair rewards, unclear instructions, punishments, and conditional positive regard. Most behaviours were drawn from self-determination theory and transformational leadership theory. We think theories like these are critical for *explaining* why these behaviours might be important, but we also argue that theories hope to *predict* the outcomes that educational psychologists care about. We describe and use robust methods for testing how well statistical models predict outcomes using data not used to build the model. If practices like these become more widespread, we hope educational psychology can move toward being able to explain *and* predict key outcomes.

Declarations

References

Abbott-Chapman, J., Martin, K., Ollington, N., Venn, A., Dwyer, T., & Gall, S. (2014). The longitudinal association of childhood school engagement with adult educational and occupational achievement: Findings from an Australian national study. *British Educational Research Journal, 40*(1), 102–120. https://doi.org/10.1002/berj.3031

Ahmadi, A., Noetel, M., Parker, P., Ryan, R. M., Ntoumanis, N., Reeve, J., Beauchamp, M., Dicke, T., Yeung, A., Ahmadi, M., Bartholomew, K., Chiu, T. K. F., Curran, T., Erturan, G., Flunger, B., Frederick, C., Froiland, J. M., González-Cutre, D., Haerens, L., Lonsdale, C. (2023). A classification system for teachers' motivational behaviors recommended in self-determination theory interventions. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000783

Andermo, S., Hallgren, M., Nguyen, T.-T.-D., Jonsson, S., Petersen, S., Friberg, M., Romqvist, A., Stubbs, B., & Elinder, L. S. (2020). School-related physical activity interventions and mental health among children: A systematic review and meta-analysis. *Sports Medicine - Open, 6*(1), 25. https://doi.org/10.1186/s40798-020-00254-x

Archambault, I., Janosz, M., Fallu, J.-S., & Pagani, L. S. (2009). Student engagement and its relationship with early high school dropout. *Journal of Adolescence, 32*(3), 651–670. https://doi.org/10.1016/j.adolescence.2008.06.007

Australian Curriculum, Assessment and Reporting Authority. (2011). *Guide to understanding ICSEA*. http://docs.acara.edu.au/resources/Guide_to_understanding_ICSEA.pdf. Accessed 30 Oct 2018.

Bailey, R. (2006). Physical education and sport in schools: A review of benefits and outcomes. *The Journal of School Health, 76*(8), 397–401. https://doi.org/10.1111/j.1746-1561.2006.00132.x

Barkoukis, V., Ntoumanis, N., & Thøgersen-Ntoumani, C. (2010). Developmental changes in achievement motivation and affect in physical education: Growth trajectories and demographic differences. *Psychology of Sport and Exercise, 11*(2), 83–90. https://doi.org/10.1016/j.psychsport.2009.04.008

Bass, B. M. (1985). Leadership: Good, better, best. *Organizational Dynamics, 13*(3), 26–40. https://doi.org/10.1016/0090-2616(85)90028-2

Bass, B. M., & Riggio, R. E. (2006). *Transformational leadership*. Psychology Press.

Beauchamp, M. R., & Morton, K. L. (2011). Transformational teaching and physical activity engagement among adolescents. *Exercise and Sport Sciences Reviews, 39*(3), 133–139. https://doi.org/10.1097/JES.0b013e31822153e7

Beauchamp, M. R., Barling, J., Li, Z., Morton, K. L., Keith, S. E., & Zumbo, B. D. (2010). Development and psychometric properties of the transformational teaching questionnaire. *Journal of Health Psychology, 15*(8), 1123–1134. https://doi.org/10.1177/1359105310364175

Beauchamp, M. R., Liu, Y., Morton, K. L., Martin, L. J., Wilson, A. H., Wilson, A. J., Sylvester, B. D., Zumbo, B. D., & Barling, J. (2014). Transformational teaching and adolescent physical activity: Multilevel and mediational effects. *International Journal of Behavioral Medicine, 21*(3), 537–546. https://doi.org/10.1007/s12529-013-9321-2

Beauchamp, M. R., Barling, J., & Morton, K. L. (2011). Transformational teaching and adolescent self-determined motivation, self-efficacy, and intentions to engage in leisure time physical activity: A randomised controlled pilot trial. *Applied Psychology. Health and Well-Being*, *3*(2), 127–150. https://onlinelibrary.wiley.com/doi/full/ https://doi.org/10.1111/j.1758-0854.2011.01048.x

Biddle, S. J., Wang, C. K. J., Kavussanu, M., & Spray, C. (2003). Correlates of achievement goal orientations in physical activity: A systematic review of research. *European Journal of Sport Science: EJSS: Official Journal of the European College of Sport Science, 3*(5), 1–20. https://doi.org/10.1080/17461390300073504

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development, 78*(1), 246–263. https://doi.org/10.1111/j.1467-8624.2007.00995.x

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Bremer, E., & Cairney, J. (2018). Fundamental movement skills and health-related outcomes: A narrative review of longitudinal and intervention studies targeting typically developing children. *American Journal of Lifestyle Medicine, 12*(2), 148–159. https://doi.org/10.1177/1559827616640196

Castillo, I., Molina-García, J., Estevan, I., Queralt, A., & Álvarez, O. (2020). Transformational teaching in physical education and students' leisure-time physical activity: The mediating role of learning

climate, passion and self-determined motivation. *International Journal of Environmental Research and Public Health*, *17*(13). https://doi.org/10.3390/ijerph17134844

Castro-Schilo, L., & Grimm, K. J. (2018). Using residualized change versus difference scores for longitudinal research. *Journal of Social and Personal Relationships, 35*(1), 32–58. https://doi.org/10.1177/0265407517718387

Cheon, S. H., Reeve, J., & Moon, I. S. (2012). Experimentally based, longitudinally designed, teacher-focused intervention to help physical education teachers be more autonomy supportive toward their students. *Journal of Sport & Exercise Psychology, 34*(3), 365–396. https://doi.org/10.1123/jsep.34.3.365

Christian, B. (2020). *The alignment problem: Machine learning and human values*. WW Norton & Company.

Cohen, R. J., Swerdlik, M., & Sturman, E. (2012). *Psychological testing and assessment - An introduction to tests & measurement: Eighth Edition*. McGraw-Hill Higher Education.

Curran, T., & Standage, M. (2017). Psychological needs and the quality of student engagement in physical education: Teachers as key facilitators. *Journal of Teaching in Physical Education: JTPE, 36*(3), 262–276. https://doi.org/10.1123/jtpe.2017-0065

Dermitzaki, I., Leondari, A., & Goudas, M. (2009). Relations between young students' strategic behaviours, domainspecific self-concept, and performance in a problem-solving situation. *Learning and Instruction, 19*(2), 144–157. https://doi.org/10.1016/j.learninstruc.2008.03.002

DeWalt, D. A., Rothrock, N., Yount, S., Stone, A. A., PROMIS Cooperative Group. (2007). Evaluation of item candidates: the PROMIS qualitative item review. *Medical Care, 45*(5 Suppl 1), S12–S21. https://doi.org/10.1097/01.mlr.0000254567.79743.e2

Digelidis, N., Papaioannou, A., Laparidis, K., & Christodoulidis, T. (2003). A one-year intervention in 7th grade physical education classes aiming to change motivational climate and attitudes towards exercise. *Psychology of Sport and Exercise, 4*(3), 195–210. https://doi.org/10.1016/S1469-0292(02)00002-X

Dinsmore, D. L., Fryer, L. K., & Dumas, D. G. (n.d.). A theoretical and metatheoretical reframing of the development of cognitive processing and learning. *Educational Psychology Review*, *this issue*.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography, 36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Dweck, C. S. (2013). *Self-theories: Their role in motivation, personality, and development*. Psychology Press.

Dweck, C. S. (2008). *Mindset: The new psychology of success*. Ballantine Books.

Elliot, A. J., Murayama, K., & Pekrun, R. (2011). A 3× 2 achievement goal model. *Journal of Educational Psychology*, *103*(3), 632. http://psycnet.apa.org/fulltext/2011-13620-001.html. Accessed 15 Oct 2018.

Elliot, A. J., & McGregor, H. A. (2001). A 2 × 2 achievement goal framework. *Journal of Personality and Social Psychology, 80*(3), 501–519. https://doi.org/10.1037/0022-3514.80.3.501

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., Qian, J., & Yang, J. (2022). *Lasso and elastic-net regularized generalized linear models* (Version 4.1–4) [Computer software]. Comprehensive R Archive Network (CRAN). https://cran.r-project.org/web/packages/glmnet/index.html. Accessed 29 June 2022.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Gable, R. A., Hester, P. H., Rock, M. L., & Hughes, K. G. (2009). Back to basics: Rules, praise, ignoring, and reprimands revisited. *Intervention in School and Clinic, 44*(4), 195–205. https://doi.org/10.1177/1053451208328831

García-Hermoso, A., Ramírez-Vélez, R., Lubans, D. R., & Izquierdo, M. (2021). Effects of physical education interventions on cognition and academic performance outcomes in children and adolescents: A systematic review and meta-analysis. *British Journal of Sports Medicine*. https://doi.org/10.1136/bjsports-2021-104112

Gumus, S., Bellibas, M. S., Esen, M., & Gumus, E. (2018). A systematic review of studies on leadership models in educational research from 1980 to 2014. *Educational Management Administration & Leadership, 46*(1), 25–48. https://doi.org/10.1177/1741143216659296

Hartwig, T. B., Sanders, T., Vasconcellos, D., Noetel, M., Parker, P. D., Lubans, D. R., Andrade, S., Ávila-García, M., Bartholomew, J., Belton, S., Brooks, N. E., Bugge, A., Cavero-Redondo, I., Christiansen, L. B., Cohen, K., Coppinger, T., Dyrstad, S., Errisuriz, V., Fairclough, S., …, Del Pozo Cruz, B. (2021). School-based interventions modestly increase physical activity and cardiorespiratory fitness but are least effective for youth who need them most: an individual participant pooled analysis of 20 controlled trials. *British Journal of Sports Medicine.* https://doi.org/10.1136/bjsports-2020-102740

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Hattie, J. (2008). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. *Routledge.* https://doi.org/10.4324/9780203887332

Hoch, J. E., Bommer, W. H., Dulebohn, J. H., & Wu, D. (2018). Do ethical, authentic, and servant leadership explain variance above and beyond transformational leadership? A meta-analysis. *Journal of Management, 44*(2), 501–529. https://doi.org/10.1177/0149206316665461

Huang, C. (2012). Discriminant and criterion-related validity of achievement goals in predicting academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(1), 48–73. https://doi.org/10.1037/a0026223

Hulleman, C. S., Schrager, S. M., Bodmann, S. M., & Harackiewicz, J. M. (2010). A meta-analytic review of achievement goal measures: Different labels for the same constructs or different constructs with similar labels? *Psychological Bulletin, 136*(3), 422–449. https://doi.org/10.1037/a0018947

Jaeger, B. C., Tierney, N. J., & Simon, N. R. (2020). When to impute? Imputation before and during cross-validation. In *arXiv [stat.ML]*. arXiv. http://arxiv.org/abs/2010.00718

Jaitner, D., Rinas, R., Becker, C., Niermann, C., Breithecker, J., & Mess, F. (2019). Supporting subject justification by educational psychology: A systematic review of achievement goal motivation in school physical education. *Frontiers in Education, 4*, 70. https://doi.org/10.3389/feduc.2019.00070

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R. *Springer.* https://doi.org/10.1007/978-1-4614-7138-7

Janosz, M., Archambault, I., Morizot, J., & Pagani, L. S. (2008). School engagement trajectories and their differential predictive relations to dropout. *The Journal of Social Issues, 64*(1), 21–40. https://doi.org/10.1111/j.1540-4560.2008.00546.x

Kanat-Maymon, Y., Shoshani, A., & Roth, G. (2021). Conditional regard in the classroom: A double-edged sword. *Frontiers in Psychology, 12*, 621046. https://doi.org/10.3389/fpsyg.2021.621046

Kirpich, A., Ainsworth, E. A., Wedow, J. M., Newman, J. R. B., Michailidis, G., & McIntyre, L. M. (2018). Variable selection in omics data: A practical evaluation of small sample sizes. *PloS One, 13*(6), e0197910. https://doi.org/10.1371/journal.pone.0197910

Kolovelonis, A., Goudas, M., & Dermitzaki, I. (2011). The effects of instructional and motivational self-talk on students' motor task performance in physical education. *Psychology of Sport and Exercise, 12*(2), 153–158. https://doi.org/10.1016/j.psychsport.2010.09.002

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, 28*, 1–26. https://doi.org/10.18637/jss.v028.i05

Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. Taylor & Francis. https://bookdown.org/max/FES. Accessed 3 Aug 2022.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., R Core Team, & et al. (2022). *Classification and regression training* (Version 6.0–92) [Computer software]. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=caret. Accessed 28 Jun 2022.

Lam, S.-F., Jimerson, S., Wong, B. P. H., Kikas, E., Shin, H., Veiga, F. H., Hatzichristou, C., Polychroni, F., Cefai, C., Negovan, V., Stanculescu, E., Yang, H., Liu, Y., Basnett, J., Duck, R., Farrell, P., Nelson, B., & Zollneritsch, J. (2014). Understanding and measuring student engagement in school: The results of an international study from 12 countries. *School Psychology Quarterly: The Official Journal of the Division of School Psychology, American Psychological Association, 29*(2), 213–232. https://doi.org/10.1037/spq0000057

Lazowski, R. A., & Hulleman, C. S. (2016). Motivation interventions in education: A meta-analytic review. *Review of Educational Research, 86*(2), 602–640. https://doi.org/10.3102/0034654315617832

Le, H., Schmidt, F. L., Harter, J. K., & Lauver, K. J. (2010). The problem of empirical redundancy of constructs in organizational research: An empirical investigation. *Organizational Behavior and Human Decision Processes, 112*(2), 112–125. https://doi.org/10.1016/j.obhdp.2010.02.003

Lei, H., Cui, Y., & Zhou, W. (2018). Relationships between student engagement and academic achievement: A meta-analysis. *Social Behavior and Personality: An International Journal, 46*(3), 517–528. https://doi.org/10.2224/sbp.7054

Linnenbrink-Garcia, L., Tyson, D. F., & Patall, E. A. (2008). When are achievement goal orientations beneficial for academic achievement? A closer look at main effects and moderating factors. *Revue internationale de psychologie sociale*, Tome 21(1), 19–70. https://www.cairn.info/revue-internationale-de-psychologie-sociale-2008-1-page-19.htm. Accessed 5 Aug 2020.

Local Burden of Disease WaSH Collaborators. (2020). Mapping geographical inequalities in access to drinking water and sanitation facilities in low-income and middle-income countries, 2000–17. *The Lancet Global Health, 8*(9), e1162–e1185. https://doi.org/10.1016/S2214-109X(20)30278-3

Marks, H. M. (2000). Student engagement in instructional activity: Patterns in the elementary, middle, and high school years. *American Educational Research Journal, 37*(1), 153–184. https://doi.org/10.3102/00028312037001153

Marmot, M. (2005). Social determinants of health inequalities. *The Lancet, 365*(9464), 1099–1104. https://doi.org/10.1016/S0140-6736(05)71146-6

Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Arens, A. K. (2019). The murky distinction between self-concept and self-efficacy: Beware of lurking jingle-jangle fallacies. *Journal of Educational Psychology, 111*(2), 331–353. https://doi.org/10.1037/edu0000281

Mâsse, L. C., O'Connor, T. M., Tu, A. W., Watts, A. W., Beauchamp, M. R., Hughes, S. O., & Baranowski, T. (2016). Are the physical activity parenting practices reported by US and Canadian parents captured in currently published instruments? *Journal of Physical Activity & Health, 13*(10), 1070–1078. https://doi.org/10.1123/jpah.2016-0012

Molnar, C. (2021). *Interpretable machine learning: A guide for making black box models explainable.* https://christophm.github.io/interpretable-ml-book/

Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological Review, 91*(3), 328–346. https://doi.org/10.1037/0033-295X.91.3.328

Noetel, M., Griffith, S., Delaney, O., Harris, N. R., Sanders, T., Parker, P., del Pozo Cruz, B., & Lonsdale, C. (2021). Multimedia design for learning: An overview of reviews with meta-meta-analysis. *Review of Educational Research*. https://doi.org/10.3102/00346543211052329

Ntoumanis, N., & Biddle, S. J. (1999). Affect and achievement goals in physical activity: A meta-analysis. *Scandinavian Journal of Medicine & Science in Sports, 9*(6), 315–332. https://doi.org/10.1111/j.1600-0838.1999.tb00253.x

O'Donoghue, G., Kennedy, A., Puggina, A., Aleksovska, K., Buck, C., Burns, C., Cardon, G., Carlin, A., Ciarapica, D., Colotto, M., Condello, G., Coppinger, T., Cortis, C., D'Haese, S., De Craemer, M., Di Blasio, A., Hansen, S., Iacoviello, L., Issartel, J., …, Boccia, S. (2018). Socio-economic determinants of physical activity across the life course: A "DEterminants of DIet and Physical ACtivity" (DEDIPAC) umbrella literature review. *PloS One, 13*(1), e0190737. https://doi.org/10.1371/journal.pone.0190737

Parker, P. D., Jerrim, J., Chmielewski, A. K., & Marsh, H. W. (2018). Predicting university entry using machine-based models and solutions. In I. Schoon & R. K. Silbereisen (Eds.), *Pathways to adulthood: Educational opportunities, motivation and attainment in times of social change* (vol. 92, pp. 92–110). UCL IOE Press.

Pearl, J. (2010). An introduction to causal inference. *The International Journal of Biostatistics, 6*(2), 7. https://doi.org/10.2202/1557-4679.1203

Peralta, L. R., Mihrshahi, S., Bellew, B., Reece, L. J., & Hardy, L. L. (2019). Influence of school-level socioeconomic status on children's physical activity, fitness, and fundamental movement skill levels. *The Journal of School Health, 89*(6), 460–467. https://doi.org/10.1111/josh.12761

Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., & Przybocki, M. A. (2021). Four principles of explainable artificial intelligence. *National Institute of Standards and Technology*. https://doi.org/10.6028/nist.ir.8312

Quin, D. (2017). Longitudinal and contextual associations between teacher–student relationships and student engagement: A systematic review. *Review of Educational Research, 87*(2), 345–387. https://doi.org/10.3102/0034654316669434

R Core Team. (2020). *R: A language and environment for statistical computing* (Version 3.6.3) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Reeve, J. (2012). A Self-determination theory perspective on student engagement. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), Handbook of Research on Student Engagement (pp. 149–172). Springer US. https://doi.org/10.1007/978-1-4614-20187_7

Rees, T., Alexander Haslam, S., Coffee, P., & Lavallee, D. (2015). A social identity approach to sport psychology: Principles, practice, and prospects. *Sports Medicine, 45*(8), 1083–1096. https://doi.org/10.1007/s40279-015-0345-4

Reeve, J., & Lee, W. (2014). Students' classroom engagement produces longitudinal changes in classroom motivation. *Journal of Educational Psychology*, *106*(2), 527. https://psycnet.apa.org/journals/edu/106/2/527.html?uid=2013-38868-001. Accessed 7 Aug 2020

Reeve, J., & Cheon, S. H. (2021). Autonomy-supportive teaching: Its malleability, benefits, and potential to improve educational practice. *Educational Psychologist, 56*(1), 54–77. https://doi.org/10.1080/00461520.2020.1862657

Reeve, J., Cheon, S. H., & Jang, H.-R. (2019). A teacher-focused intervention to enhance students' classroom engagement. In *Handbook of student engagement interventions* (pp. 87–102). Elsevier. https://www.sciencedirect.com/science/article/pii/B9780128134139000073

Reschly, A. L., & Christenson, S. L. (2012). Jingle, jangle, and conceptual haziness: Evolution and future directions of the engagement construct. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of Research on Student Engagement* (pp. 3–19). Springer US. https://doi.org/10.1007/978-1-4614-2018-7_1

Roorda, D. L., Koomen, H. M. Y., Spilt, J. L., & Oort, F. J. (2011). The influence of affective teacher–student relationships on students' school engagement and achievement: A meta-analytic approach. *Review of Educational Research, 81*(4), 493–529. https://doi.org/10.3102/0034654311421793

Roorda, D. L., Jak, S., Zee, M., Oort, F. J., & Koomen, H. M. Y. (2017). Affective teacher–student relationships and students' engagement and achievement: A meta-analytic update and test of the mediating role of engagement. *School Psychology Review, 46*(3), 239–261. https://doi.org/10.17105/SPR-2017-0035.V46-3

Rosenzweig, E. Q., & Wigfield, A. (2016). STEM motivation interventions for adolescents: A promising start, but further to go. *Educational Psychologist, 51*(2), 146–163. https://doi.org/10.1080/00461520.2016.1154792

Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, 101860. https://doi.org/10.1016/j.cedpsych.2020.101860

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *The American Psychologist, 55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Ryan, R. M., & Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. The Guilford Press.

Scheinost, D., Noble, S., Horien, C., Greene, A. S., Lake, E. M., Salehi, M., Gao, S., Shen, X., O'Connor, D., Barron, D. S., Yip, S. W., Rosenberg, M. D., & Constable, R. T. (2019). Ten simple rules for predictive modeling of individual differences in neuroimaging. *NeuroImage, 193*, 35–45. https://doi.org/10.1016/j.neuroimage.2019.02.057

Selvitopu, A., & Kaya, M. (2021). A meta-analytic review of the effect of socioeconomic status on academic performance. *Journal of Education*, 00220574211031978. https://doi.org/10.1177/00220574211031978

Shmueli, G. (2010). To explain or to predict? *Statist. Sci., 25*(3), 289–310. https://doi.org/10.1214/10-STS330

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., …, Nosek, B. A. (2018). Many analysts, one data set: making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L., & Macnamara, B. N. (2018). To what extent and under which circumstances are growth mind-sets important to academic achievement? Two meta-analyses. *Psychological Science, 29*(4), 549–571. https://doi.org/10.1177/0956797617739704

Skinner, E. A., & Pitzer, J. R. (2012). Developmental dynamics of student engagement, coping, and everyday resilience. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of Research on Student Engagement* (pp. 21–44). Springer US. https://doi.org/10.1007/978-1-4614-2018-7_2

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology, 85*(4), 571–581. https://doi.org/10.1037/0022-0663.85.4.571

Slavich, G. M., & Zimbardo, P. G. (2012). Transformational teaching: Theoretical underpinnings, basic principles, and core methods. *Educational Psychology Review, 24*(4), 569–608. https://doi.org/10.1007/s10648-012-9199-6

Smith, N., Quested, E., Appleton, P. R., & Duda, J. L. (2016). A review of observational instruments to assess the motivational environment in sport and physical education settings. *International Review of Sport and Exercise Psychology, 9*(1), 134–159. https://doi.org/10.1080/1750984X.2015.1132334

Sosu, E. M., Dare, S., Goodfellow, C., & Klein, M. (2021). Socioeconomic status and school absenteeism: A systematic review and narrative synthesis. *Review of Education*, *9*(3). https://doi.org/10.1002/rev3.3291

Su, Y.-L., & Reeve, J. (2011). A meta-analysis of the effectiveness of intervention programs designed to support autonomy. *Educational Psychology Review, 23*(1), 159–188. https://doi.org/10.1007/s10648-010-9142-7

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Teixeira, P. J., Carraca, E. V., Markland, D., Silva, M. N., & Ryan, R. M. (2012). Exercise, physical activity, and self-determination theory: A systematic review. *The International Journal of Behavioral Nutrition and Physical Activity, 9*, 78. https://doi.org/10.1186/1479-5868-9-78

Tomaszewski, W., Xiang, N., & Western, M. (2020). Student engagement as a mediator of the effects of socioeconomic status on academic performance among secondary school students in Australia. *British Educational Research Journal, 46*(3), 610–630. https://doi.org/10.1002/berj.3599

Torchiano, M. (2020). *Efficient effect size computation: Package "effsize"* (Version 0.8.1) [R]. https://cran.r-project.org/web/packages/effsize/effsize.pdf. Accessed 5 Aug 2022.

Tsai, Y.-M., Kunter, M., Lüdtke, O., Trautwein, U., & Ryan, R. M. (2008). What makes lessons interesting? The role of situational and individual factors in three school subjects. *Journal of Educational Psychology, 100*(2), 460–472. https://doi.org/10.1037/0022-0663.100.2.460

Turnnidge, J., & Côté, J. (2018). Applying transformational leadership theory to coaching research in youth sport: A systematic literature review. *International Journal of Sport and Exercise Psychology, 16*(3), 327–342. https://doi.org/10.1080/1612197X.2016.1189948

Tutz, G., & Ramzan, S. (2015). Improved methods for the imputation of missing data by nearest neighbor methods. *Computational Statistics & Data Analysis, 90*, 84–99. https://doi.org/10.1016/j.csda.2015.04.009

Usami, S., Murayama, K., & Hamaker, E. L. (2019). A unified framework of longitudinal models to examine reciprocal relations. *Psychological Methods, 24*(5), 637–657. https://doi.org/10.1037/met0000210

Van Yperen, N. W., Blaga, M., & Postmes, T. (2014). A meta-analysis of self-reported achievement goals and nonself-report performance across three achievement domains (work, sports, and education). *PloS One, 9*(4), e93594. https://doi.org/10.1371/journal.pone.0093594

Vasconcellos, D., Parker, P. D., Hilland, T., Cinelli, R., Owen, K. B., Kapsal, N., Lee, J., Antczak, D., Ntoumanis, N., Ryan, R. M., & Lonsdale, C. (2019). Self-determination theory applied to physical education: A systematic review and meta-analysis. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000420

Vella, S. A., Braithewaite, R. E., Gardner, L. A., & Spray, C. M. (2016). A systematic review and meta-analysis of implicit theory research in sport, physical activity, and physical education. *International Review of Sport and Exercise Psychology, 9*(1), 191–214. https://doi.org/10.1080/1750984X.2016.1160418

Walters, G. D. (2016). Beyond Dustbowl Empiricism: The need for theory in recidivism prediction research and its potential realization in causal mediation analysis. *Criminal Justice and Behavior*, *44*(1). https://doi.org/10.1177/0093854816677566

White, R. L., Bennie, A., Vasconcellos, D., Cinelli, R., Hilland, T., Owen, K. B., & Lonsdale, C. (2021). Self-determination theory in physical education: A systematic review of qualitative studies. *Teaching and Teacher Education, 99*, 103247. https://doi.org/10.1016/j.tate.2020.103247

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement, 20*(1), 59–69. https://doi.org/10.1177/014662169602000106

Wilson, A. J., Liu, Y., Keith, S. E., Wilson, A. H., Kermer, L. E., Zumbo, B. D., & Beauchamp, M. R. (2012). Transformational teaching and child psychological needs satisfaction, motivation, and engagement in elementary school physical education. *Sport, Exercise, and Performance Psychology, 1*(4), 215–230. https://doi.org/10.1037/a0028635

Woods, C. B., Tannehill, D., & Walsh, J. (2012). An examination of the relationship between enjoyment, physical education, physical activity and health in Irish adolescents. *Irish Educational Studies, 31*(3), 263–280. https://doi.org/10.1080/03323315.2012.710068

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B, Statistical Methodology, 67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Authors and Affiliations

Michael Noetel[1] · Philip Parker[2] · Theresa Dicke[2] · Mark R. Beauchamp[3] · Nikos Ntoumanis[4,5] · Ryan M. Hulteen[6] · Carmel Diezmann[7] · Alexander Yeung[2] · Asghar Ahmadi[2] · Diego Vasconcellos[2] · John Mahoney[8] · Poulomee Datta[9] · Scott Doidge[10] · Chris Lonsdale[2]

✉ Chris Lonsdale
chris.lonsdale@acu.edu.au

1   The University of Queensland, Brisbane, Australia

2   Institute for Positive Psychology and Education, Australian Catholic University, Sydney, Australia

3   University of British Columbia, Vancouver, Canada

4   University of Southern Denmark, Odense, Denmark

5   Halmstad University, Halmstad, Sweden

6   Louisiana State University, Baton Rouge, USA

7   Queensland University of Technology, Brisbane, Australia

8   School of Health and Behavioural Sciences, Australian Catholic University, Brisbane, Australia

9   Macquarie University, Sydney, Australia

10   Australian Catholic University, Melbourne, Australia