Use of automated coding methods to assess motivational behaviour in education

Ahmadi, Asghar

Ahmadi, Asghar. (2022). Use of automated coding methods to assess motivational behaviour in education [PhD Thesis]. Australian Catholic University https://doi.org/10.26199/acu.8yw1v

Use of Automated Coding Methods to Assess Motivational Behaviour in Education

Submitted by

Asghar Ahmadi

BA. (Secondary Education - Education)

MA. (Physical Education and Sport Sciences - Sport Psychology)

A thesis submitted in total fulfilment of the requirements for the degree of

Doctor of Philosophy

Institute for Positive Psychology and Education

Faculty of Health Sciences

Australian Catholic University

14 July 2022

Principal Supervisor: Dr Michael Noetel

Co-Supervisor: Professor Chris Lonsdale

Co-Supervisor: Professor Philip Parker

Statement of Authorship

This thesis contains no material published elsewhere or extracted in whole or in part from a thesis by which I have qualified for or been awarded another degree or diploma. No parts of this thesis have been submitted towards the award of any other degree or diploma in any other tertiary institution. No other person's work has been used without due acknowledgment in the main text of the thesis. All research procedures reported in the thesis received the approval of the relevant Ethics/Safety Committees (where required).

[redacted]

Asghar Ahmadi

14 July 2022

Statement of Appreciation

The completion of this thesis was an enormous endeavour and I could not have accomplished it without the support of my supervisors and encouragement from my colleagues, family, and friends.

First and foremost, I am thankful to my supervisors, Mike, Chris, and Phil. Saying thank you doesn't seem to be enough now that you've helped me complete this journey. You were an integral part of me being able to make it to the end. You provided invaluable academic advice in an excellent manner while being open-hearted and patient. Your ongoing support, immense knowledge and experience have encouraged and motivated me to persist and learn more throughout my PhD. I cannot be thankful enough. Thank you so much.

Mike, throughout my PhD, you fully supported me by providing invaluable and constructive feedback and being available when I needed you. I liked the way you introduced extra resources for learning and provided step-by-step feedback to better develop the required skills. The other skills I learned from you have been helpful for me beyond my academic life. Your support, help, and care has been exceptional. Thank you so much for your extensive support during the completion of my research studies and thesis.

Chris, thank you for your outstanding supervision, mentorship and patience that cannot be underestimated. You have been an extraordinary mentor for me during my PhD and even before I come to Australia. You encouraged, supported, challenged, and promoted me during my PhD. Thank you for assisting with practical, technical and theoretical aspects of my research. Your expertise, guidance and encouragement have been beyond valuable for my thesis and my life. Thank you so much for providing me with the required facilities and various learning, professional and enriching experiences.

And Phil, I appreciate your academic guidance, good manner, and calmness throughout my PhD. You provided technical feedback and insightful comments on my

studies which have been invaluable. You helped me to be a more skilled and knowledgeable researcher with a better understanding of the professional and societal context. It was inspiring to see you always with a smile on your face, and a positive mood. I will use your selfless support as a model as I move forward.

I would like to thank all the members of the Institute for Positive Psychology and Education that I had the opportunity to work with. I am extremely grateful to Prof. Ryan and Prof. Reeve for their invaluable feedback on my research and for their support, time and consideration. Thank you to the IPPE members for all the jokes, laughs, and fun activities. Thank you for always being willing to help. To Devan, for his help with PhD studies, teaching, and random games. To Borja, for his support, jokes, and catch-ups. And to Prof. Marsh and Joy for the refreshing games of table tennis.

I am thankful to my research studies' co-authors, including the systematic review, Delphi study and dictionary study. I am tremendously grateful for their time, collaboration, and valuable input to the research.

To my friends near and far, I could not get to this phase without your ongoing support, help and hope. I have always felt your encouragement and great support. I am lucky to have you and I am really thankful.

And special thanks to my parents and family, your continual love and support has been amazing. You always supported my autonomy, helped me and motivated me to pursue my goals. I want to thank my mother and father for their ongoing love, support and for their dedication to my education since my childhood. I am so grateful to my siblings, for their unconditional love.

Table of Contents

List of Tables

List of Figures

Abstract

Teachers' motivational behaviour is related to important student outcomes. Assessing teachers' motivational behaviour has been helpful to improve teaching quality and enhance student outcomes. However, researchers in educational psychology have relied on self-report or observer ratings. These methods face limitations on accurately and reliably assessing teachers' motivational behaviour; thus restricting the pace and scale of conducting research. One potential method to overcome these restrictions is automated coding methods. These methods are capable of analysing behaviour at a large scale with less time and at low costs. In this thesis, I conducted three studies to examine the applications of an automated coding method to assess teacher motivational behaviours. First, I systematically reviewed the applications of automated coding methods used to analyse helping professionals' interpersonal interactions using their verbal behaviour. The findings showed that automated coding methods were used in psychotherapy to predict the codes of a well-developed behavioural coding measure, in medical settings to predict conversation patterns or topics, and in education to predict simple concepts, such as the number of open/closed questions or class activity type (e.g., group work or teacher lecturing). In certain circumstances, these models achieved near human level performance. However, few studies adhered to best-practice machine learning guidelines. Second, I developed a dictionary of teachers' motivational phrases and used it to automatically assess teachers' motivating and de-motivating behaviours. Results showed that the dictionary ratings of teacher need support achieved a strong correlation with observer ratings of need support ($r_{full\ dictionary}$ = .73). Third, I developed a classification of teachers' motivational behaviour that would enable more advanced automated coding of teacher behaviours at each utterance level. In this study, I created a classification that includes 57 teacher motivating and de-motivating behaviours that are consistent with self-determination theory. Automatically assessing teachers' motivational

behaviour with automatic coding methods can provide accurate, fast pace, and large scale analysis of teacher motivational behaviour. This could allow for immediate feedback and also development of theoretical frameworks. The findings in this thesis can lead to the improvement of student motivation and other consequent student outcomes.

Chapter 1 | Introduction and Overview

Introduction

Teacher behaviours are the main components of the teaching process and play an important role in student outcomes. Particularly, teachers' motivational behaviour is a key determinant of student motivation, engagement, and achievement (Korpershoek et al., 2016; Lazowski & Hulleman, 2016; Reeve, 2009; Reeve & Cheon, 2021; Reeve & Jang, 2006; Ryan & Deci, 2017; Vasconcellos et al., 2020). By engaging in certain behaviours, teachers can foster high quality motivation (i.e., autonomous motivation) in their students. Students with such teachers may benefit from positive behavioural, cognitive, and affective outcomes (Bartholomew et al., 2018; Jang et al., 2010; Reeve et al., 2004; Tessier et al., 2010b). On the other hand, teachers may behave in a way that hinders high quality motivation and enjoyment. Students in such classes might experience maladaptive outcomes such as ill-being, burnout, depression and negative affect (Bartholomew et al., 2011; Vansteenkiste & Ryan, 2013). That is, teachers' behaviour has a significant impact on student motivation and outcomes. Fortunately, teachers can learn how to avoid detrimental behaviours and use more adaptive behaviours (Reeve & Cheon, 2021; Su & Reeve, 2011). As a result, researchers have designed interventions grounded in theories to help teachers to become more motivating (Lazowski & Hulleman, 2016; Reeve & Cheon, 2021). Such interventions have been efficient and helpful, and have successfully changed teacher behaviours to become more supportive of students psychological needs (Reeve & Cheon, 2021). To provide helpful interventions, researchers must assess teacher behaviours before and after an intervention. However, researchers usually relied on traditional methods such as self-report or observer coding of teacher behaviours (Smith et al., 2016). These methods are expensive and prone to bias and inaccuracies (Haerens et al., 2013a; Kahneman et al., 2021). One possible solution to these limitations are automated methods of coding teacher behaviour. In this thesis, I aimed to explore automated coding methods of assessing teachers' motivational behaviour.

Teacher Behaviour and Motivation Theories

Teachers' motivational behaviour is a key determinant of the quality of students' motivation (Reeve, 2009; Reeve & Cheon, 2021; Reeve & Jang, 2006; Vasconcellos et al., 2019). These behaviours have been addressed via different motivation theories. For example, Achievement Goal Theory (AGT) classifies teacher-created motivational climates as mastery (or 'task-involving'; where a teacher focuses on learners' self-referenced improvement and effort) or performance ('ego-involving'; where they focus on learners' competence compared to others, Ames, 1992; Dweck, 1999). Research findings have shown that more mastery-oriented climates were related to higher enjoyment, satisfaction and engagement (Liukkonen et al., 2010; Kaplan & Maehr, 2007; Standage & Treasure, 2002; Treasure & Robert, 2001). More performance-oriented climates were related to negative or maladaptive consequences such as lower levels of intrinsic motivation and engagement (Duda et al., 2014; Standage & Treasure, 2002).

Teachers can create less performance-oriented, and more mastery-oriented climates in a range of ways (Duda et al., 2014). For example, Ames (1992) introduced the TARGET model of building a mastery-climate. It suggests teachers can increase mastery orientation by changing the:

Task: by including variety, challenge, and purpose in the learning activities;

Authority: by fostering active participation and sense of ownership;

Recognition: by focusing on individual progress and improvement;

Grouping: by using individual and cooperative learning;

Evaluation: by using diverse methods to assess and monitor learning; and

Time: by allowing students to participate in scheduling and sequencing.

Meta-analyses of Achievement Goal Theory find that the orientations students hold are strong predictors of engagement and persistence. However, interventions built on

Achievement Goal Theory are only moderately effective, compared to interventions underpinned by other theories (Lazowski & Hulleman, 2016). This may be due to the fact that Achievement Goal Theory focuses on only one of three basic psychological needs. A mastery climate is designed to build a more sustainable, less fragile sense of competence, but focuses less on students' connections with others (relatedness) or feeling of self-direction (autonomy).

In contrast, Self-Determination Theory is a broad set of propositions that suggests a range of influential teacher motivational behaviours. Based on this theory, teachers' interpersonal behaviours can support or thwart all three basic psychological needs (competence, autonomy, and relatedness; Ryan & Deci, 2017). For example, teachers can apply a need supportive style by asking the student's perspective, inviting input, and energising the student's own motivational resources. In a need thwarting climate, students might be told what to do or pressured for compliance (Reeve & Cheon, 2021).

Many studies have shown that a need supportive climate is associated with positive student outcomes such as enjoyment and engagement (Ntoumanis & Standage, 2009; Standage et al., 2007). Need thwarting climates are associated with maladaptive outcomes such as ill-being, burnout, and negative affect (Bartholomew et al., 2011; Vansteenkiste & Ryan, 2013). Meta-analyses of interventions underpinned by self-determination theory also suggest those interventions are among the best for improving motivation (Lazowski & Hulleman, 2016). Motivated students are more engaged in classroom activities and achieve better academic outcomes (Froiland & Oros, 2014; Froiland & Worrell, 2016; Gottfried et al., 2008; Reeve, 2009; Vansteenkiste et al., 2008). On average, student motivation declines over time, and teacher behaviour plays a moderating role (Gillet et al., 2012; Gnambs & Hanfstingl, 2016; Lepper et al., 2005). That is, some teachers accelerate this decline and others can reverse the trend. But, research has shown that teachers can learn to adopt need supportive styles to better motivate students and prevent this decline. For this reason, a substantial proportion of SDT-based studies applied interventions to help teachers to be more motivating.

Researchers in education mostly use observational or self-report methods to assess teacher motivational behaviour (Cheon et al., 2012; Van den Berghe et al., 2016). These methods hinder the pace of research as they need large amounts of time and financial resources. One method to overcome those limitations is via automated coding methods. For example, the dictionary method is an automated coding method that can replicate manual coding in a timely and efficient way (Nelson et al., 2018). This method has been used to assess family adjustments and conflicts (Robbins et al., 2013), depression and depression-vulnerability (Rude et al., 2004), and health (Eichstaedt et al., 2015). It works via searching for the presence of a set of predetermined words (i.e., a dictionary word) that represents a psychological construct. This has some important limitations. For example, it does not account for the context (e.g., words used before or after a dictionary word). Regardless, it is useful because it has been shown to be a successful tool in extracting various psychological constructs without the requirements of manual coding (Tausczik & Pennebaker, 2010).

More sophisticated methods of automated coding exist. For example, in psychotherapy, Tanana et al. (2016) used automated coding methods to annotate the specific technique used in each sentence from a psychologist. By being able to meaningfully code each sentence, these methods allow for more fine-grained feedback. Rather than merely describing an overall ‘gestalt’ of whether the therapist was ‘supportive’ (for example), it allows for feedback on specifically where the therapist was supportive (e.g., the therapist reflected the client’s emotions in these 7 instances). To do this, the authors used support vector machines, which are sophisticated machine learning models that encode meaning and allow for dynamic interactions between words (e.g., “it sounds like” being modelled as 'reflective listening'). However, to code such a model, Tanana et al. (2016) required 175,000 of coded utterances. That is, humans needed to listen to 341 therapy sessions and, for each sentence, describe what the therapist was doing (e.g., 'reflective listening').

Annotated datasets this large are not available in education. As a result, I plan to test a dictionary method using currently available data to see if they are viable methods of automatically coding teacher behaviour. In my systematic review, I plan to find all studies of automated coding methods from helping professions to see what datasets are available, and what methods from other fields might be applicable in education. In addition, one advantage for automated coding in psychotherapy is that there exist well-established coding frameworks for annotating transcripts (e.g., the 'motivational interviewing skills code'). This is not available in education. So, in my Delphi study, I plan to develop a coding framework for use in education. Overall, through this thesis, I aimed to develop better methods for automatically coding teacher motivational behaviour.

In the field of education, some social cognition and motivational theories have been used to explain motivation in students (e.g., Transformational Leadership Theory, Achievement Goal Theory, Implicit Theories of Intelligence, and Self-Determination Theory). Among the theories explaining motivation in education, Self-Determination Theory (SDT) has been an established theoretical framework to investigate student motivation (Lazowski & Hulleman, 2016; Reeve & Cheon, 2021; Vasconcellos et al., 2020). It is among the most well researched motivation theories (Lazowski & Hulleman, 2016). This theory focuses on some key determinants of behaviour (e.g., psychological need satisfaction) and how those determinants influence both the quality and quantity of motivation (Hagger et al., 2020, p. 104). Further, SDT highlights the inner motivational resources that all students have (e.g., intrinsic and extrinsic values), as well as suggestions on how teachers might involve, support and vitalise these resources during the teaching process to facilitate high-quality student engagement (Niemiec & Ryan, 2009; Reeve, 2012). This theory has been an effective framework for examining and enhancing student motivation and consequent student outcomes (Niemiec & Ryan, 2009; Vasconcellos et al., 2020).

SDT outlines propositions that are explicit, detailed, and dynamic, and were supported among a variety of nations, cultures and contexts (Legault, 2017). Lazowski and Hulleman's (2016) meta-analysis of interventions in the educational context indicated that those based on the SDT framework have a greater effect than Achievement Goal Theory and Implicit Theories of Ability (Cohen's d for SDT = 0.70, 95% CI[0.53, 0.78], Cohen's d for AGT = 0.38, 95% CI[0.09, 0.67], Cohen's d for ITA = 0.56, 95% CI[0.31, 0.80]). SDT focuses on the interpersonal style in the teaching process between a teacher and students, an essential component of educational environments (Ryan & Deci, 2017). My second and third studies aim to assess teachers' interpersonal behaviour that influences student motivation. In the second chapter, I synthesised studies that applied an automated coding method to assess interpersonal interactions in helping professionals. In the third study, I aimed to test the applications of an automated coding method in assessing teachers' motivational behaviour based on SDT. Therefore, in this thesis, I focus on self-determination theory due to its focus on teachers' interpersonal style in motivating students towards engagement and achievement; however, a brief introduction to other theories related to student motivation is warranted.

Transformational Leadership Theory

Transformational Leadership Theory (Burns, 1978) is a theory of leadership that explains how instructors can create a positive change in their subjects. The concept of Transformational Leadership was first introduced by James MacGregor Burns in the political leadership context, but it is now used in diverse business, government and educational domains. This theory involves a process in which "leaders and their followers raise one another to higher levels of morality and motivation" (Burns, 1978, p. 20). This process is performed through four behavioural dimensions including idealised influence, inspirational motivation, intellectual stimulation, and individualised consideration. *Idealised Influence* is characterised by displays of integrity, acting as ideal role model for followers and embodying the qualities that the leader wants from the followers. *Inspirational Motivation* refers to the degree to which the leader inspires and motivates the followers to achieve their goals. *Intellectual Stimulation* takes place when a leader challenges the current assumptions, encourages followers to see the issues from multiple aspects and fosters independent thoughts. Finally, *Individualized Consideration* refers to the extent to which the leader responds to the individuals' needs and builds ever-stronger, trust-based relationships with the followers.

Transformational Leadership Theory has been informed by a considerable body of research in a variety of contexts including business, military, sports, and education (for a review, see Wang et al., 2011). The results have shown that transformational leadership is consistently associated with positive employee or follower outcomes such as enhanced performance (Judge & Piccolo, 2004), motivation (Piccolo & Colquitt, 2006), and well-being (Arnold et al., 2007). In the educational context, the aim of previous transformational leadership research was to study the influence of school principals' behaviour on teachers (e.g., Bogler, 2001; Ross & Gray, 2006). In recent years, research mostly done by Beauchamp and his colleagues extended the conceptual frameworks of transformational leadership to provide foundations to understand the effects of PE teachers transformational behaviour on student outcomes (Beauchamp et al., 2010, e.g., 2011). As a rationale for this extension, they assert that in both teaching and leadership environments the focus is on influencing others to achieve their goals (Beauchamp & Morton, 2011). In the educational settings, students are considered analogous to "employees" in the work settings.

Transformational Teaching Structure. Transformational teaching refers to applying transformational theory tenets in the teaching context. Beauchamp et al. (2011; 2014) suggested that to foster optimal motivation, teachers could display 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). A number of studies have indicated that displays of transformational teaching structures are related to cognitive, affective and behavioural student consequences such as self-determined motivation, self-efficacy, in-class engagement, and stronger intentions for leisure-time engagement in physical activity (Beauchamp et al., 2011, 2014; Morton et al., 2010).

Achievement Goal Theory

Achievement Goal Theory (Ames, 1992; Dweck, 1999; Nicholls, 1984, 1989; Roberts, 2001) is a social cognitive theory of motivation and provides insight into the quality and quantity of motivation in achievement settings. AGT posits that the primary motive of individuals striving in achievement contexts is to demonstrate competence or achievement. Based on AGT, there are two different goal states that reflect how individuals construe their success and improvement. These two major states are *task* and *ego* involved goals (Nicholls, 1984). Within a *task*-involved goal, success is defined as working hard to master the skills, achieve the optimal performance, and the perceived competence is self-referenced. In an *ego*-involved goal, other-referenced criteria are considered the definition of success. In such goals, success is experienced when an individual outperforms others, wins the match or is superior to others, therefore, competence is defined in terms of interpersonal or normative comparisons.

AGT postulates that there are two concepts that impact the task and/or ego involvement of a goal: *achievement goal orientations* and *achievement goal climate* (Ames, 1992; Dweck & Leggett, 1988). *Achievement goal orientations* reflect individuals' stable differences associated with goals within a specific context. Two dispositional tendencies were identified as "task orientation" and "ego orientation". A task-oriented person tends to compare themselves with self-referenced criteria and self-improvement, whereas an ego-oriented person tends to define success as outperforming peers and displaying superior ability. These two goal orientations determine different consequences in an achievement context (Chazan et al., 2022). In the field of education and particularly physical education, research findings showed that task-orientation is associated with positive and adaptive outcomes such as higher levels of intrinsic motivation, enjoyment, satisfaction and engagement (Kaplan & Maehr, 2007; Standage & Treasure, 2002; Treasure & Robert, 2001). On the other hand, ego-orientation is related to maladaptive outcomes such as lower levels of intrinsic motivation and engagement (Standage & Treasure, 2002).

Achievement goal climate refers to the motivational climate created by significant others, such as teachers or coaches (Ames, 1992). The goal climate can be task- (or mastery-) focused and/or ego-(or performance-) focused (Ames, 1992). In a task-focused climate, the main focus is on task mastery and self-improvement while in an ego focused climate focus is on normative competence and outperforming others. When teachers create a motivational climate, they lead students to generally adopt a task- or ego-involved goals (Ames, 1992). Previous studies have shown a link between task-oriented motivational climate with positive outcomes such as intrinsic motivation, intention for participation in later physical activities and belief that motivation or effort caused success and satisfaction (Escartí & Gutiérrez, 2001; Hein et al., 2004; Treasure & Robert, 2001). In contrast, an ego-oriented climate has a negative effect on enjoyment and students' preference for challenging tasks (Dweck & Leggett, 1988; Treasure & Robert, 2001).

Implicit Theories of Intelligence

A student's Implicit Theory of Intelligence (Dweck, 1999; Dweck & Leggett, 1988) reflects their often unconscious (hence implicit) beliefs about the malleability of their abilities. Some students may believe that ability and intelligence are fixed and difficult to change (an *entity* theory or belief). Others may perceive that such attributes can be developed by effort (an incremental theory or growth mindset; Blackwell et al., 2007). Individuals who hold an entity belief, tend to believe that abilities are stable, and their performance is a consequence of that stability about their abilities (Dweck, 1999). They consider their ability as a "thing" that might have a set quantity. On the contrary, individuals with a growth mindset believe that the ability can be developed with effort and through time. These individuals tend to participate in challenging tasks and activities that improve their skills and abilities (Dweck, 1999).

Implicit theories have been widely examined in educational contexts both in observational and intervention designs (Sisk et al., 2018; Warburton & Spray, 2017). Studies have shown that in comparison with entity beliefs, growth mindsets are associated with higher levels of motivation towards physical education and sport and enjoyment of physical activity (Biddle et al., 2003), lower levels of anxiety (Ommundsen, 2001), higher academic engagement and mastery-oriented strategies (Burnette et al., 2013; Robins & Pals, 2002), and higher academic engagement (Martin et al., 2013). However, pooled effects appear to be small (Sisk et al., 2018) to moderate (Lazowski & Hulleman, 2016). Nevertheless, over the long-term it may be beneficial for students to adopt growth mindsets (Funder & Ozer, 2019), and so teachers may benefit their students by praising students for their effort and improvement, rather than students' innate talents and skills.

Self-Determination Theory

Self-Determination Theory (SDT; Deci & Ryan, 1985) is a theoretical framework that has been applied to explain human behaviour and motivation in various contexts. This meta-theory provides insights into how and why people are participating in an activity by focusing on personal and situational factors. Specifically, it focuses on the role of social, environmental and individual factors that promote self-motivation and psychological adjustment. In the last two decades, SDT has been intensively applied in education (Guay, 2022; Howard et al., 2021; Reeve & Cheon, 2021). Currently, this theory consists of six mini-theories, each of them addressing one aspect of motivation or functioning.

Cognitive Evaluation Theory. Cognitive Evaluation Theory (Deci & Ryan, 1985; Ryan, 1991) is the first mini-theory of SDT that exclusively addresses the effects of external factors on intrinsic motivation. Cognitive Evaluation Theory asserts that external conditions, such as rewards and praise, can enhance or undermine intrinsic motivation. CET includes three main propositions (Ryan & Deci, 2017). First, it proposes that extrinsic factors relevant to the initiation of behaviour will affect individuals' intrinsic motivation through the changes in the locus of causality and perceived autonomy. Locus of causality refers to the motivation of intentional behaviour and it can be internal or external. Autonomy refers to the extent that individuals perceive that their behaviour is self-endorsed and self-directed in an activity. Based on CET, events will decrease intrinsic motivation if they promote an external locus of causality or thwart perceived autonomy. On the contrary, events will enhance intrinsic motivation if they promote internal locus of causality and satisfy autonomy.

Second, CET proposes that external events will affect intrinsic motivation to the extent that they influence perceived competence (Ryan & Deci, 2017). External factors that promote greater perceived competence will increase intrinsic motivation, while events that frustrate perceived competence, diminish intrinsic motivation (Ryan & Deci, 2017). Third, CET asserts that any external event related to motivation can have three functional aspects: informational, controlling, and amotivating. The informational aspect fosters an internal perceived locus of causality and perceived competence and positively influences intrinsic motivation. For example, a teacher might praise a student because of the student's progress. The controlling aspect facilitates external locus of causality (a person's perception of the cause of success or failure) by undermining perceived autonomy and competence and thus, negatively influences intrinsic motivation. For example, a teacher might pressure students to engage in certain activities to merely achieve higher scores. And finally, the amotivating aspect facilitates perceived incompetence by undermining both intrinsic and extrinsic motivation and promotes amotivated functioning. CET asserts that the relative salience of the three aspects of an event to a student determines its effect on intrinsic motivation.

A large body of research has examined the tenets of CET. For instance, Goudas and colleagues (1994) examined CET propositions in high school PE classes, and showed that perceived autonomy and competence predicts intrinsic motivation . A meta-analysis examining the effect of extrinsic rewards on intrinsic motivation showed that external events, such as positive verbal feedback, significantly increases intrinsic motivation, while tangible rewards might decrease competence and intrinsic motivation (Vansteenkiste & Deci, 2003). Based on CET, when the informational aspect of positive verbal feedback is more than its controlling aspect, then its informative aspect is more salient to students. A systematic review on the effect of praise on intrinsic motivation showed that praise with an informational aspect promotes intrinsic motivation while a controlling aspect undermines it (Henderlong & Lepper, 2002; Vasconcellos et al., 2020). In total, research findings provided strong support for CET in experimental and real-life educational settings.

Organismic Integration Theory. Organismic Integration Theory (OIT) was developed to expand Cognitive Evaluation Theory by addressing the concept of extrinsic motivation in its various forms. This mini-theory outlines that students are sometimes extrinsically motivated to participate in activities. That is, students are more motivated when they participate in an activity to get a reward or benefit, and less motivated if they engage in an activity to avoid an unpleasant situation such as punishment. OIT asserts that there are different forms of extrinsic motivation. Further, it explains the developmental process of internalisation and integration of extrinsically motivated behaviour (Ryan & Deci, 2017). Based on OIT, there are four different types of extrinsic motivation. *External regulation* refers to performing an activity in order to "satisfy an external demand or reward contingency" (Ryan & Deci, 2000, p. 72). *Introjected regulation* refers to doing an activity to feel proud of doing that activity or to avoid feeling guilty (Ryan & Deci, 2000, p. 72). With *identified regulation*, a behaviour is performed because individuals value the behaviour and perceive it to be consistent with their goals (Ryan & Deci, 2000, p. 72). Finally, *integrated regulation* occurs when individuals perceive the behaviour as an integral part of who they are and the behaviour is in congruence with individuals' sense of self (Deci & Ryan, 2002).

Identified, integrated and intrinsic motivation represent 'autonomous' or 'self-determined' motivation (Ryan & Deci, 2017). These forms of motivation are most frequently associated with student positive outcomes such as student success and well-being, persistence and performance goals (Howard et al., 2021; Vasconcellos et al., 2020). On the other hand, external and introjected regulations are considered 'controlled' or 'non-self-determined' motivation and are associated with maladaptive outcomes such as decreased well-being (Howard et al., 2021; Vasconcellos et al., 2020).

OIT also postulates that there is another motivation regulation, namely amotivation, which stands separate from extrinsic motivation. Amotivation refers to a state in which a person faces either lack of intention for participating in an action or a lack of interest in the outcomes it might yield. Research findings showed that higher levels of self-determined motivation are associated with greater enjoyment (Ntoumanis, 2002), intention to be physically active and effort (Ntoumanis, 2001) and leisure-time physical activity (Hagger et al., 2003), while non-self-determined motivation is a strong predictor of boredom (Ntoumanis, 2002) and feelings of unhappiness (a systematic review; Howard et al., 2021; Standage et al., 2005).

Causality Orientations Theory (COT). While CET and OIT focus on the influence of the social environment on individuals' motivation, Causality Orientation Theory (COT) considers the stable individual differences towards the factors that affect motivation. In other words, COT assumes that there are significant differences among individuals' perceptions of the social context (so-called causality orientations) as autonomy supportive or controlling (Ryan & Deci, 2017). Generally, COT posits that there are three distinct kinds of these orientations; autonomy, controlled, and impersonal orientations. Autonomy-oriented individuals tend to regulate behaviour as self-originated and volitional. These individuals tend to perceive external incentives as informational and supportive of their self-determination. In contrast, controlled oriented individuals tend to look toward pressures and controls in the environment to perceive the behaviour as originating from outside the self. Finally, impersonal causality of orientation refers to the lack of control over causality which leads to amotivation. COT claims that an autonomous orientation is most related to positive motivation and health outcomes, while controlled and impersonal orientations are not associated with such outcomes (Ryan & Deci, 2017).

COT posits that these causality orientations are developmental but can be influenced by biological and social environmental factors over time. Also, it proposes that the intensity of all three orientations may differ from person to person, but everybody inherently possesses these orientations (Deci & Ryan, 2002). For example, each student in a class is somewhat oriented to both self-improvement (autonomy-orientated) and the pressures of academic grades (controll-orientated). Still, some are more focused on one of these approaches than the other. A considerable body of research examined COT propositions in various contexts. Results showed that autonomy orientation is associated with autonomous forms of motivation and various adaptive outcomes such as self-esteem, well-being and engagement in daily activities (Deci & Ryan, 1985; Hagger et al., 2015; Weinstein et al., 2012). On the other hand, research findings indicated that controlled and impersonal orientations are associated with external motivation and maladaptive outcomes such as daily stress and self-consciousness (Deci & Ryan, 1985; Ryan & Deci, 2017).

Basic Psychological Needs Theory (BPNT). Basic Psychological Needs Theory (Deci & Ryan, 2000; Ryan & Deci, 2017) postulates that human beings are born with three innate basic psychological needs which are universal: *Autonomy*, *Competence*, and *Relatedness*. *Autonomy* is defined as the extent that individuals perceive themselves as the origin of their decisions and behave with a sense of volition and willingness. *Competence* refers to the degree to which individuals feel confident and effective in action. Finally, *Relatedness* refers to the feelings of being connected to others (e.g., peers, parents, coaches) and belonging to groups in social environments (Deci & Ryan, 2002). SDT proposes that the impacts of the social environment on optimal functioning are not direct. Instead, it is mediated by the satisfaction or frustration of the psychological needs. It postulates that when teachers satisfy these needs, students will become more autonomously motivated and function better. In contrast, when these needs are thwarted, students will become less motivated, engaged, and perform less well (Ryan & Deci, 2000).

A large body of research has shown a clear empirical link between psychological needs satisfaction with behavioural, cognitive, and psychological consequences in a variety of domains including sports (Adie et al., 2008; Vlachopoulos et al., 2013), parenting (Costa et al., 2015), work (Olafsen et al., 2018; Slemp et al., 2018), education (Aelterman et al., 2016;

Howard et al., 2021; Sheldon et al., 2009), and physical education (Haerens et al., 2015; Ntoumanis & Standage, 2009; Standage et al., 2005). For example, Standage and colleagues (2005) examined the impact of environment and basic needs satisfaction on student outcomes. Their findings showed that satisfaction of needs predicts intrinsic motivation and persistence on challenging tasks. Also, research findings indicate that satisfaction of basic psychological needs predicts various cognitive, affective and behavioural outcomes such as psychological growth, preference for challenging tasks, and well-being (e.g., see two reviews: Ntoumanis & Standage, 2009; Standage et al., 2007). In comparison with needs dissatisfaction, experiences of need frustration are strongly associated with maladaptive outcomes such as ill-being, disordered eating, burnout, depression and negative affect (Bartholomew et al., 2011; Vansteenkiste & Ryan, 2013).

Need Supportive vs. Need Thwarting Style. According to BPNT, social and environmental factors such as school environment or significant others (e.g., teachers, peers, parents) can influence psychological needs (Ryan & Deci, 2017). Based on SDT, teachers in educational settings can adopt a needs supportive or needs thwarting style in relation to the students' basic needs (Reeve & Cheon, 2021). In a *Need Supportive style*, the three basic needs can be nurtured by providing rationales and choices (Autonomy Supportive; Mageau et al., 2015), acknowledging individuals' progress and providing positive feedback (Competence Supportive; Sheldon & Filak, 2008) or providing interpersonal closeness and respect (Relatedness Supportive; Vansteenkiste et al., 2010). On the other hand, in a *Need Thwarting* style, the basic needs are thwarted by using controlling language or external rewards (Autonomy Thwarting; K. JBartholomew et al., 2009), demeaning individuals' ability or emphasising their failures (Competence Thwarting; Sheldon & Filak, 2008), or putting them in an isolated state or neglecting them (Relatedness Thwarting; Sheldon & Filak, 2008; Vansteenkiste et al., 2010). Many researchers have studied the relationship between teaching style and student outcomes (Reeve & Cheon, 2021). For instance, Haerens et al. (2013) examined the mediating role of needs satisfaction between teacher behaviour and student outcomes. Results showed that a needs supportive style led to adaptive student outcomes such as autonomous motivation and engagement. In contrast, a needs thwarting style led to negative outcomes, such as amotivation and disengagement, via need frustration (Van den Berghe et al., 2016).

Goal Contents Theory (GCT). The fifth mini-theory, Goal Contents Theory (GCT) was created to explain the goals of participating in an activity by addressing questions such as "what is the student's goal of engaging in a class?" Particularly, GCT asserts that different goal contents affect motivation and well-being differently (Vansteenkiste et al., 2006). Intrinsic goals are inherently valued goals such as personal growth, deeper interpersonal relationships, and physical health. For example, a student's goal of participating in physical activity might be intrinsic goals such as achieving physical health or performance progress. These goals are associated with intrinsic motivation and well-being and other positive outcomes (Vansteenkiste et al., 2006). GCT postulates that this effect takes place through the satisfaction of basic needs: that is, these goals foster students' basic needs, which consequently will enhance student outcomes. In contrast, extrinsic goals refer to instrumental outcomes such as enhanced status, fame, or appearance. Extrinsic goals of a student might be teacher satisfaction or showing off to the class. These goals may thwart psychological needs because they often impede basic needs. They are therefore associated with lower intrinsic motivation and greater ill-being (Kasser & Ryan, 1996). Another GCT proposition is that success at attaining intrinsic goals could yield enhanced wellness while attaining extrinsic goals might tend to be related to less enhanced wellness (Bradshaw et al., 2021).

GCT's hypotheses have been supported across different cultures and nations. For example, Grouzet and his colleagues (2005) examined the tenets of GCT among 15 cultures. Their findings showed that intrinsic goals such as affiliation and physical health are related to each other, and different from extrinsic goals (Grouzet et al., 2005). Also, previous correlational studies found that basic needs satisfaction and motivation mediate the relationship between intrinsic and extrinsic goals and well-being (Gunnell et al., 2014).

The important indication of GCT is that achieving success in pursuit of extrinsic goals undermines psychological needs and well-being (Niemiec et al., 2009; Reeve, 2012; Vansteenkiste et al., 2008). On the other hand, the pursuit of intrinsic goals leads to positive psychological and behavioural outcomes such as performance, psychological well-being, engagement and higher persistence in learning activities (Vansteenkiste et al., 2004, 2006). To sum up, GCT underlies the relationship between the intrinsic and extrinsic goals and motivation, and indicates that performing a behaviour in pursuit of extrinsic goals can be detrimental.

Relationships Motivation Theory (RMT). The sixth and most recent mini-theory of SDT is Relationships Motivation Theory (RMT). According to RMT, high-quality relationships satisfy an individuals' need for autonomy, relatedness and competence. RMT proposes that relationships with others are not only desirable for people, but essential for motivation and well-being (Ryan & Deci, 2017). Furthermore, RMT asserts that while the need for relatedness drives the initial desire to experience and maintain close relationships, high-quality relationships should not just satisfy the need for relatedness, but all three psychological needs (Ryan & Deci, 2017). For example, La Guardia et al. (2000) showed that meaningful relationships satisfy the needs for relatedness, and autonomy (and to a lesser degree: competence).

Autonomy Supportive Versus Controlling Environment. According to SDT, teachers' behaviour in terms of interpersonal communication may be viewed as autonomy supportive or controlling. A controlling environment (e.g., using controlling language or pressure) facilitates an external perceived locus of causality, reduces perceived autonomy, and the corresponding extrinsic motivation. On the other hand, an autonomy-supportive environment (e.g., providing rationale and choice, acknowledging individuals feelings and perceptions) facilitates an intrinsic locus of causality. It also satisfies the need for autonomy, leading to intrinsic motivation. It has been argued that when students' increasing desire for autonomy is met with a more controlling atmosphere, maladaptive outcomes such as a decline in engagement occurs (Creasey & Jarvis, 2012). In contrast, research has shown that autonomy supportive environments created by teachers are associated with more adaptive outcomes such as intrinsic motivation, engagement and behavioural persistence (Bryan & Solmon, 2007; Standage et al., 2006).

SDT and its mini-theories have great potential to explain the reasons that facilitate or hinder student motivation in the educational context and also consequent student outcomes. In summary, CET explains how teachers' interpersonal style affects students' intrinsic motivation, and OIT explains the extrinsic motives of participation and how an individual may internalise the values and regulations associated with behaviours. COT explains how some students can see the same teacher behaviour as controlling while others see the same behaviour as autonomy supportive. BPNT focuses on the effects of satisfying or thwarting the innate motivational resources in the class environment. GCT explains why pursuing some goals—even when successful—can lead to lower wellbeing. And finally, RMT describes the importance of developing and maintaining meaningful relationships with others.

Measuring Teacher Behaviour in SDT

Numerous SDT-based studies have examined teachers' motivational behaviour and its impact on students' motivation and other outcomes (Cheon et al., 2012; Franco & Coterón, 2017; Guay, 2022). These studies heavily relied on the traditional methods of assessing behaviour. For example, many studies used self-report questionnaires (Cheon et al., 2012; Cox & Williams, 2008; Hagger et al., 2003; Marsh et al., 2006). Also, observational methods have been used in the research in motivation including live observation (Jang et al., 2010; Reeve et al., 2004) and video rating by human raters (Sarrazin et al., 2006; Van den Berghe et al., 2016). Furthermore, an integrated approach of both self-report and observational methods have been used (Boyce et al., 2009; Mahoney et al., 2016; Morgan et al., 2005). Even though self-report methods have yielded many important findings, they are prone to bias, and observational methods are extremely costly (Moyers, Martin, et al., 2005). Accurate coding of data, inter-rater and intra-rater reliability are issues that could also limit the applications of observational methods (Atkins et al., 2012). Observer coding methods require coders to observe a full-length session. Furthermore, with the advances in scalable intervention approaches, researchers can conduct large scale studies and collect huge amounts of observational data. Consequently, coding such datasets requires large human and financial resources. These limitations restrict applications of the traditional methods to analyse large volumes of data (Imel et al., 2015). Accurate assessment of motivational behaviour is important since bias in the assessment will lead to misleading conclusions, but cost-effective methods are also required to manage large amounts of available data. To approach this aim, in the second chapter, I conducted a systematic review of automated methods. And in the third chapter, I tested the efficacy of an automated method to assess TMBs. In the following, I present previously applied methods used to assess teacher motivational behaviours and the need for new methods.

Motivational Behaviour Assessment in Educational Context

Researchers mainly relied on traditional methods, such as self-report or observation, to assess motivational behaviour in education. More recently, new methods of behaviour assessment have been used in other contexts, such as motivational interviewing. These methods are rarely used to assess teachers' motivational behaviour, however, they have great potential to be applied in the educational context because the motivational constructs largely overlap.

Traditional Methods

Self-Report Instruments. Self-report instruments, such as questionnaires, are an easy and fast method that can be applied in small- and large-scale studies. In education, most of the researchers assessing motivational behaviour and climate used self-report instruments (Bartholomew et al., 2011; Chatzisarantis & Hagger, 2009; Cheon et al., 2012; Cox & Williams, 2008; Hagger et al., 2009). Even though there are pragmatic benefits of self-report tools, they face shortcomings. In a class, it has been shown that there is a significant difference between teachers' self-reports of their behaviour and observer ratings of their behaviour (Haerens et al., 2013b; Smith et al., 2015; Van den Berghe et al., 2013). For example, Haerens and colleagues found no significant relationship between the observer-rated and student-perceived need support (Haerens et al., 2013b). They suggested that this difference might stem from the influence of "key moments" on subjects or "general perceptions" rather than specific perceptions of the climate (Haerens et al., 2013b; Smith et al., 2015). Also, 'social desirability' can affect self-report data, in which individuals tend to present themselves in a way that is more socially acceptable (Fisher, 1993). Similarly, a 'halo effect' can mean that irrelevant characteristics, like physical attractiveness, can influence students' ratings of their teachers (Riniolo et al., 2006). Different students may interpret items in different ways, and some students may have poor memories of pertinent details, making some reports unreliable (Van den Berghe et al., 2016). To overcome these restrictions, researchers used observational methods to assess motivational behaviour in the classroom environment.

Observational Methods. Throughout educational psychology, observational methods have been used for assessing teacher behaviours. These methods have been widely applied to collect data about motivational climate (Haerens et al., 2013b; Morgan et al., 2005; Sarrazin et al., 2006; Tessier et al., 2010a; Van den Berghe et al., 2013, 2016). Numerous authors have argued that these methods lead to more reliable and valid results than self-report (Morgan et al., 2005; Smith et al., 2016; Van den Berghe et al., 2013). Applying these methods reduces individual differences in question interpretation (Haerens et al., 2013b). Also, the observational methods make it possible to rate the teachers' behaviour on an individual level (e.g., at a sentence level) or as a whole over a class (Haerens et al., 2015; Smith et al., 2016). Coders can also be trained in theoretical models so they know what to look for with more astute judgement. For example, observers can be trained to rate 'autonomy support' generally where students must rate specific behaviours hypothesised to relate to the construct.

The observational method has been used in a variety of approaches to assess teacher behaviours. In earlier attempts, observers conducted live observations (Jang et al., 2010; e.g., Reeve et al., 2004). In this method, the observer (or rater) is present in the class and scores the behaviour as it occurs by using pre-defined checklists. This method is helpful to solve the perception and interpretation bias; however, it might disturb the natural ongoing process of the class because of the Hawthorne effect (or observer effect). Also, it might be too difficult to replicate this procedure for reliability purposes. Further, the rater might not be able to consider all aspects of the behaviour simultaneously (Latvala et al., 2000). To prevent this, videotaped observations have been used as an alternative method to live observation. Using video recording of behaviour, most teacher behaviours are recorded faithfully, therefore making it possible to replicate rating without the need for multiple observers being present in the classroom. Also, videotaped behaviour can be analysed in different ways by focusing on the specific aspects of behaviour or climate (Latvala et al., 2000). For example, a videotaped class might be used to analyse teachers' motivational behaviour in one study and the same recording might be used to analyse student engagement level in another. And finally, because of the high ecological validity of the videotape approach, the desired behaviour can be rated based on the duration, frequency or intensity (Haerens et al., 2013b).

There are some limitations with observational methods. Observational methods still remain very expensive, time-consuming, and labour intensive. This method requires training coders, conducting the observation, and the coding process itself (Atkins et al., 2012; Moyers, Martin, et al., 2005). Although observational methods reduce variability in question interpretation, low interrater reliability between observers is still common (Haerens et al., 2013b). Observers' beliefs and tendencies might also affect coding. Furthermore, at least two coders are needed to code the same observation to assess inter-rater reliability. This makes observational methods even more expensive. New methods such as dictionaries and machine learning might be efficient methods to replace manual coding, with considerably less time and financial cost. So, we aimed to use a dictionary to test the applicability of this method in assessing TMBs quickly and cost efficiently.

Automated Coding Methods and Machine Learning

Researchers have used automated coding approaches to code various desired behaviours of helping professionals. As I will outline in my Study 1, researchers in psychotherapy have applied automated coding systems to code counsellor-client interactions (Ahmadi et al., 2021). For example, these methods have been applied to evaluate provider fidelity for motivational interviewing (i.e., the extent that a counsellor performed in accordance with the recommended instructions; Atkins et al., 2014). These methods are a good option because they are time, cost, and human resource-efficient. For example, applying automated coding systems saves on the costs of training and paying coders. Applying automated coding systems on a larger scale only requires additional computational costs which are insignificant in comparison with the huge costs of manual coding (Can et al., 2016). Moreover, it can solve coder reliability and coder drift concerns (changes in coders' implicit definitions of the target behaviour over time; Bakeman & Quera, 2011, p. 23). So, these methods hold promise to accelerate and simplify the procedure of behaviour coding. For this reason, in my thesis, I reviewed the implementation of automated methods in assessing interpersonal behaviour in chapter 2, and tested the efficiency of an automated coding method in assessing teachers' motivational behaviour in chapter 3.

Automated coding methods can process large datasets of behaviour at significantly lower costs compared to traditional methods (Ahmadi et al., 2021). These methods, under some circumstances, can reach high correlations with human coders (Ahmadi et al., 2021). One of the commonly applied automated coding methods includes Natural Language Processing (NLP). NLP takes the input data such as transcriptions of language to extract data and mainly involves preprocessing and classification techniques. Preprocessing involves taking the transcript and formatting it in a way that is easier to automatically classify. For example, stemming is a pre-processing step where grammatical suffixes are removed (e.g., -ed, -ing, -ation) and only the stem remains (e.g., car, cars, car's, cars' become car). Preprocessing steps like stemming are useful because most of the meaning can be preserved while making the modelling task simpler. Classification techniques then involve trying to add labels to the text, such as the sentiment of the text (e.g., 'positive vs. negative'; Sentiment Analysis) or the topic of the text (e.g., 'liberal vs. conservative issues'; Topic Modeling). As mentioned earlier, there are many different methods of classifying text. One interpretable method that uses NLP is the dictionary method, which counts the number of words that appear in each dataset that have been assigned a particular category. In this thesis, I used the dictionary method to examine the implementation of automated assessment of motivational behaviour in education.

One set of automated coding methods for analysing teacher behaviour is machine learning. Machine Learning is defined as a "field of study that gives computers the ability to learn without being explicitly programmed" (Simon, 2013, p. 13). It is a branch of artificial intelligence based on the idea that systems can learn from input data, identify patterns and make decisions with minimal human intervention. Data scientists use many different kinds of machine learning algorithms to discover patterns in big data sets that lead to actionable insights. These different algorithms can be classified into two main groups based on the way they “learn” from data to make predictions: supervised and unsupervised learning (described below; James et al., 2013). There are also some other methods such as semi-supervised learning. Overall results from the previous studies suggest that applying machine learning methods can be as efficient as manual methods. For example, Tanana and colleagues (2016) examined the performance of two automated coding models and showed that these models have ‘good-to-strong’ agreement with human coders (Cohen’s kappa > 0.60). Below I discuss supervised and unsupervised machine learning, the strengths and limitations of these approaches, and an alternative method of automatic annotation.

Supervised Machine Learning. Supervised machine learning is widely used in research and industry (Grimmer & Stewart, 2013; Jurka et al., 2012). In this method, a set of existing data is manually coded. Then, this dataset is used to train a model to code the remaining data. The model learns from the characteristics of the coded data to classify or rate the un-coded data. The process of learning from the training dataset can be thought of as a teacher supervising the model algorithm. Through methods like classification, regression, prediction, and gradient boosting, supervised learning uses patterns to anticipate the values of the label on new, unseen and unlabeled data (Mikut & Reischl, 2011).

Supervised learning models can be grouped into classification or regression, based on the task required of applying. Classification is used when the output feature is a category, such as “disease” or “no disease”. When there are only two labels, this is called binary classification. When there are more than two categories, it is called multi-class classification. In regression tasks, the output feature is a continuous value, such as “age” or “weight”. Ordinary least squares regression is one form of supervised machine learning for a regression task. Similarly, logistic regression is a commonly used machine learning model for classification. However, more complex machine learning models allow for sophisticated non-linear relationships and interactions between thousands of features. The aim of both regression and classification models is to make predictions about the future based on past data; the difference is that the dependent feature is numerical for regression and categorical for classification.

Unsupervised Machine Learning. Unsupervised machine learning models learn from the test data that have not been labelled or classified. This method identifies the categories and similar topics based on the presence or absence of commonalities in each unique piece of data. Compared with supervised learning where training data are labelled with the appropriate classifications, unsupervised models must learn the relationships between elements in a data set and classify the raw data without help. Principal component analysis is a form of unsupervised machine learning that many psychology researchers are familiar with, but again more complex methods can allow for sophisticated relationships between features. These attempts to capture the relationships can take many different algorithmic forms; however, all the models have the same approach of mimicking human logic by looking for indirect hidden structures, patterns or features to analyse the new data. The techniques in unsupervised learning include clustering and association learning. In clustering, the data is divided into several groups and the aim is to segment data into several clusters and perform analysis of each data set to find patterns. “K-mean clustering” and “Hierarchical Clustering” algorithms are the most popular and widely used algorithms for clustering in unsupervised models. Association learning, the other technique, reduces the number of features being considered to find the exact information required.

Limitations of Using Machine Learning to Assess Teacher Motivational Behaviours. Machine learning can be incredibly powerful, but training machine learning algorithms generally requires millions of coded examples of the behaviour under investigation (Brown et al., 2020; Tanana et al., 2016). Also, the machine learning models usually perform better when they are used to predict the codes of a well developed behavioural coding measure (Ahmadi et al., 2021). However, in education, there are limitations that prevent researchers using machine learning models to analyse teachers' motivational behaviour. First, a huge amount of data is seldom available in education, so few large machine learning models exist to assess teacher behaviour (Ahmadi et al., 2021). Further, a consistent behavioural measure of teachers' motivational behaviour has not yet been developed; many studies use their own measure of teacher behaviour. This thesis aims to overcome a number of these limitations by testing a dictionary as a method of automatically annotating teacher behaviour, and by laying the groundwork for more sophisticated models by creating an expert-derived classification system.

Dictionary Method. The dictionary method is a convenient tool in automated text analysis and has widely been used to replicate or replace manual coding (Nelson et al., 2018). This method provides a simple and quick way to code large volumes of text or transcribed data. In short, dictionaries can be used to extract constructs by identifying the key words used within a specific category. Tausczik and Pennebaker (2010) reviewed a range of studies and demonstrated that this method is a reliable and valid way of assessing specific concepts such as emotion and thinking styles. Research has shown that word use predicts many other concepts such as family adjustments and conflicts (Robbins et al., 2013), depression and depression-vulnerability (Rude et al., 2004), and physical health (Eichstaedt et al., 2015).

Therefore, it is reasonable to expect that dictionaries could extract motivational constructs from transcripts of teacher lessons.

A wide range of studies have developed programs to measure a variety of concepts using dictionaries (e.g., LIWC; Pennebaker et al., 2001),; (OL; Liu, 2010), (EmoLex; Mohammad & Turney, 2013). Some researchers might use pre-existing dictionaries across contexts, but they often fail to translate well across contexts because the specific words that indicate a construct is context specific (Grimmer & Stewart, 2013). For example, the word "sensitive" can be a positive adjective in one context while having a negative meaning in another. Because of these small but significant differences, it is often useful for researchers to develop dictionaries to assess their exact construct of interest, in their context of interest. Most modern tools for using dictionaries (e.g., LIWC) allow for these custom dictionaries to be added to the list of built-in dictionaries.

Linguistic Inquiry and Word Count Software (LIWC)

Linguistic inquiry and word count software (LIWC) have been used to extract various constructs in a variety of contexts including emotional expression (Bantum & Owen, 2009), mindfulness (Collins et al., 2009), lyrics content (Czechowski et al., 2016; Falk, 2013; Petterson, 2008) and, beliefs about privacy (Vasalou et al., 2011). For a review of studies which have used LIWC, see Tausczik and Pennebaker (2010). This software is extensively validated and the output results are easy to interpret (Fast et al., 2017). LIWC contains predefined and human validated dictionaries offering a wide range of vocabularies for categories such as "positive emotions", "negative emotions" and "anxiety". LIWC uses dictionaries to classify the text based on the predefined categories. Each category includes keywords specific to that category. The software checks each word of the corpus (e.g., teacher transcripts) against the words in the various dictionaries and counts the number of words corresponding to each category (e.g., anxiety). For example, suppose that the goal is to determine the extent of “negative emotions” and “positive emotions” expressed in a text. A dictionary could consist of two categories, one representing “positive emotions” and the other “negative emotions”. Each category might contain a set of keywords or stemmed words (e.g.,“happ*” that stands for happy, happiness, happiest, etc.). The LIWC program would first search each word in a corpus to find out whether that word is specified in a category or not. Then, if the word is specified in a category, that will be counted as a score for that category, otherwise, it will be labelled as “unlabelled”. This process will be done for all of the individual words in the text.

Advantages and disadvantages of using dictionaries

Dictionary methods have some advantages over traditional coding methods. First, you do not need a large dataset of annotated examples to develop a dictionary. As a result, it requires less time, budget, and resources compared with machine learning methods. Once a dictionary has been developed, it can be applied to a huge amount of data without any extra costs. They are also easy for models to be transparently tested and interpreted by other researchers, where large language models may be very powerful but famously inscrutable (Christian, 2020). Dictionaries have a number of limitations—such as their inability to account for context or complex word formations— however until educational psychology has a large set of annotated examples using a consistent classification of motivational behaviours, dictionaries may be the best method for automatically coding teacher behaviours.

Research Objectives

Despite the increasing interest in the use of automated coding methods to analyse interpersonal interactions, a systematic review of their applications across different contexts has not been done yet. Such a review would allow researchers to identify which models work well and in which contexts. Similarly, research on the accuracy and validity of these methods across contexts is scarce: automated coding methods might perform well under some circumstances but not others. However, these circumstances have not been systematically identified for the use of automated methods to assess interpersonal behaviours. Furthermore, there are well-established best practice guidelines for the use of machine learning methods. However, a synthesis of studies on the extent of adherence to the best practice guidelines is not available, meaning we do not know how confident we should be in the findings of existing research. Therefore, the aim of the first study (Chapter 2) is to address these gaps by conducting a systematic review of automated coding methods to analyse verbal behaviour in interpersonal interactions.

While the machine learning models can overcome the limitations of the traditional methods, I identified two main barriers with their applications to assess teacher motivational behaviour. First, these models rely on huge annotated datasets of behaviour. Second, to get such a big dataset, transcripts usually need to be annotated at the sentence or utterance level (rather than coded for the whole lesson, for example). However, such a huge dataset of teacher behaviours was unavailable and not feasible for this thesis. Instead, in Chapter 3, I surveyed experts in self-determination theory to create a dictionary of need-supportive and need-thwarting teaching. Using coded lessons where observers had rated teacher need-support, as a whole, I compared the performance of this dictionary against the coding of two observers.

In the third study, I created a classification system of teachers' motivational behaviour that would allow researchers to code teacher behaviours at a fine-grained level (e.g., sentence by sentence). This project aimed to serve many purposes (e.g., to promote reporting and replicability of educational interventions), but one key goal was to prepare the platform for machine learning projects in the future. While the dictionary in my second study could provide an overall estimate of teachers' motivational language within a class, dictionaries treat the whole lesson as a 'bag of words' and ignore the context. More advanced machine learning models could empirically derive the phrases that indicate need support, but require sentence-level coding to be well trained. The classification I developed aims to help future researchers reliably code teacher behaviour at each sentence level. Together, the three studies aim to inform how educational psychology researchers can better automate the assessment of teacher behaviour, improving the efficiency of research, and opening doors for practitioners to get more reliable, valid, and rapid feedback on their teaching practice.

Chapter 2 | A Systematic Review of Machine Learning for Assessment and Feedback of Treatment Fidelity

Preface

This chapter has been published (Ahmadi et al., 2021) in Psychosocial Intervention (IF = 4.58 - Q1, SJR = 0.88). I was the first author on the publication and contributed the majority (60%) of the work (see Research Portfolio Appendix). I have retained most of the language and text as published. I made some minor text changes for the context of this thesis for tables, figures, and references to appendices rather than online only supplementary material.

Abstract

Background

Many psychological treatments have been shown to be cost-effective and efficacious, as long as they are implemented faithfully. Assessing fidelity and providing feedback is expensive and time-consuming. Machine learning has been used to assess treatment fidelity, but the reliability and generalisability is unclear. We collated and critiqued all implementations of machine learning to assess the verbal behaviour of all helping professionals, with particular emphasis on treatment fidelity for therapists.

Methods

We conducted searches using nine electronic databases for automated approaches of coding verbal behaviour in therapy and similar contexts. We completed screening, extraction, and quality assessment in duplicate.

Results

Fifty-two studies met our inclusion criteria (65.3% in psychotherapy). Automated coding methods performed better than chance, and some methods showed near human-level performance; performance tended to be better with larger data sets, a smaller number of codes, conceptually simple codes, and when predicting session-level ratings than utterance-level ones. Few studies adhered to best-practice machine learning guidelines.

Conclusion

Machine learning demonstrated promising results, particularly where there are large, annotated datasets and a modest number of concrete features to code. These methods are novel, cost-effective, scalable ways of assessing fidelity and providing therapists with individualised, prompt, and objective feedback.

Keywords

Machine learning, treatment fidelity, treatment integrity, clinical supervision, feedback

Introduction

When implemented faithfully, psychological treatments are powerful (Barth et al., 2013; Blanck et al., 2018; Kazdin, 2017; Öst & Ollendick, 2017). But, a major problem with both researching and implementing psychological treatments is fidelity (Bellg et al., 2004; Perepletchikova & Kazdin, 2005). Ensuring that treatments are implemented faithfully is important for a few reasons. First, when training practitioners on evidence-based interventions, prompt clinician feedback can facilitate skill acquisition and faithful implementation (Prowse et al., 2015; Prowse & Nagel, 2015). Second, without assessing fidelity, we cannot determine whether effects from intervention studies are due to a homogenous treatment (Prowse et al., 2015; Prowse & Nagel, 2015). However, treatment fidelity is rarely well assessed; fewer than 10% of studies adequately assess fidelity (Perepletchikova et al., 2007; Perepletchikova & Kazdin, 2005). Cost and time are significant barriers (Borrelli, 2011). In psychotherapy, technology has become a well-established method of reducing costs of treatment by creating, for example, online interventions (Fairburn & Patel, 2017; Kazdin, 2017). But, the use of technologies for assessment and training is comparatively nascent (Fairburn & Cooper, 2011; Fairburn & Patel, 2017). This paper presents a systematic review of machine learning strategies to assess the fidelity of psychological treatments.

Fidelity encompasses three core components: adherence, differentiation, and competence (Rodriguez-Quintana & Lewis, 2018). Adherence describes the therapist's use of methods proposed by the guiding framework (e.g., using cognitive defusion while delivering Acceptance and Commitment Therapy). Differentiation is the avoidance of methods not proposed by that theory (e.g., using thought stopping while delivering Acceptance and Commitment Therapy). Competence is the skill with which the therapist implements the intervention (e.g., demonstrating a strong therapeutic alliance; Kazantzis, 2003). As a result, treatment fidelity is important both in the content and the process of therapy. Many interventions, like Motivational Interviewing and Cognitive Behaviour Therapy, both prescribe the content of therapy (e.g., change-talk and cognitive challenging, respectively) and the process of therapy (e.g., both emphasise the importance of an empathic therapeutic alliance; Kazantzis, 2003; Madson et al., 2009). From a content perspective, it is common for therapists to drift away from the core, evidence-based foci of therapy (Bellg et al., 2004; Waller, 2009; Waller & Turner, 2016). They may fail to use interventions that faithfully incorporate the therapy (low adherence) or 'dabble' in interventions from other therapies (low differentiation). But fidelity can also refer to the non-judgemental, compassionate, empathic process that is central to many therapies. As such, quality interpersonal interactions are critical for competent treatment (Kornhaber et al., 2016). Psychologists that competently demonstrate evidence-based interpersonal skills are more effective at reducing maladaptive behaviours such as substance abuse and risky behaviours than clinicians with poorer skills (e.g., Parsons et al., 2005). Their clients are more likely to complete treatment and change behaviour too (Golin et al., 2002; Moyers, Miller, et al., 2005; Street et al., 2009).

As a result, researchers have developed a range of treatment integrity measures (Rodriguez-Quintana & Lewis, 2018), including many that assess the content of therapy (McGlinchey & Dobson, 2003) and the process of therapy (e.g., Motivational Interviewing Skill Code: Miller et al., 2003; Motivational Interviewing Treatment Integrity: Moyers, Martin, et al., 2005). There are even measures for assessing how well treatment fidelity is assessed (Perepletchikova et al., 2009). These measures improve the quality of research and the translation of evidence-based therapies into practice (Prowse et al., 2015; Prowse & Nagel, 2015). The most objective of these measures involve an observer rating the behaviours of the therapist at regular intervals or after having watched an entire session with a client. As a result, assessing fidelity requires significant resources (Fairburn & Cooper, 2011).

Recently, researchers have begun applying machine learning models to automate this task. These models will not be useful if they fail to accurately assess fidelity, or if the methods used to create the models do not generalise to other samples. So, in this paper, we aimed to identify, synthesise, and critique the automated coding methods that have been applied to treatment fidelity.

Machine Learning. Machine learning refers to any algorithm that learns patterns from data. A linear regression model, familiar to most readers, is a form of machine learning, where an algorithm discerns the strength of the linear relationship between variables. However, machine learning also includes a broad range of other, often more complex, algorithms. These algorithms can either learn the patterns automatically by themselves (i.e., unsupervised machine learning) by, for example, identifying how data points cluster together. Alternatively, they can be trained using labelled data (i.e., supervised machine learning), where, for example, thousands of sentences are labelled by humans as 'empathic' and the model identifies the words that might indicate empathy. The line between 'statistics' and 'machine learning' is imprecise. In common usage, 'statistics' refers to more interpretable models that allow for inferences that explain a phenomenon (Hastie et al., 2009; Shmueli, 2010). 'Machine learning' is a more encompassing, umbrella term that also includes less interpretable models that may predict but not explain (Hastie et al., 2009; Shmueli, 2010). So while traditional statistics aim to explain relationships between variables, machine learning also includes methods that focus on predictive accuracy over hypothesis-driven inference (Breiman 2001). With new computational capabilities, machine learning can use large, multidimensional data to construct complex, non-linear models (Breiman et al., 2001). Traditional statistical methods are more interpretable but those constraints mean they perform less well in these more complex problems (Bi et al., 2019). This is an important feature because predicting interpersonal interactions requires multidimensional models that account for the complexity of human language.

Concept of Accuracy in Machine Learning. In machine learning, accuracy evaluates how well the model identifies relationships and patterns between variables in a dataset. Several evaluation metrics and validation methods have been used to evaluate the prediction performance and generalization of machine learning methods. The commonly used metrics include accuracy, precision, F1, sensitivity, specificity, and area under the receiver operating characteristic (AUC ROC) curve (for a description of the performance metrics, see Appendix A.1). There has been extensive debate on what metric is best for which task (Handelman et al., 2019). However, one way to choose the most appropriate metric is to consider the distribution of classes and the potential cost of misclassification (Hernandez-Orallo, 2012). For example, in psychotherapy, accuracy might be a good indication of a model's performance which shows the correct prediction of true positives out of all the observations. However, in detecting suicidality, the recall (or sensitivity) metric may be important as the correct identification of all high-risk cases may be crucial. So, considering the intended purpose of using machine learning models can be helpful to determine the most appropriate performance metric and threshold.

One of the important goals of developing ML models is to predict the outputs in the future unseen data. Validation techniques evaluate the generalizability of models to 'out of sample' data (i.e., data not used to train the model). After training a model, validation usually involves testing the model on new data that was not used in training. This is different from the common practice of looking at, for example, R-squared from the output of a regression model. Here the prediction metric—r-squared—comes from the same data used to build the model. From the perspective of machine learning, only predictive accuracy from new data—that is data not used in building the model—is of interest. In machine learning, new data is referred to as unseen data because the model has not seen the data and thus does not have the option to update the model or its parameters in response to it. Several methods have been used to validate models such as cross validation and hold-out 'train and test'. Cross-validation (which is also called internal validation) is a commonly used method where a dataset is separated into a training subset and a testing subset. Then, the prediction metrics are calculated to assess the prediction accuracy on the testing subset. Some of the cross-validation methods include split-half (50% training, 50% test samples), imbalanced-split (i.e., 70:30), k-fold (split into k subsets, usually 5 or 10), leave-one-out (a single test case is held-out of the training sample), or bootstrapping methods (Delgadillo 2021; Rodriguez et al. 2010). Another validation method, named hold-out 'train and test', better estimates the generalisability of models to future datasets. This process is called external validation, where the model is trained on some data (training dataset) and is tested on data from a different sample, study, or setting. This method is stronger than cross-validation because the validation set is more likely to be representative of future data and less likely to overlap with the training set.

Machine Learning May Improve Feedback for Therapists. Therapists vary greatly in their effectiveness, and with more experience, they actually decrease their effectiveness (Goldberg et al., 2016). This decline in effectiveness may be partially explained by lapses in fidelity. For example, without feedback or coaching, fidelity to motivational interviewing substantially decreases within six months of training (Schwalbe et al., 2014). This is often described as 'therapist drift', where well-meaning therapists fail to adhere to the prescribed practice guidelines (Waller, 2009; Waller & Turner, 2016). Therapists are bad at identifying these problems themselves because they rely on unreliable signals of their own effectiveness (Tracey et al., 2014). However, it is possible to mitigate these problems through quality feedback, auditing and supervision (Barwick et al., 2012; Ivers et al., 2012; Madson et al., 2009). Indeed, one of the core goals of training and clinical supervision is increasing treatment fidelity (Bellg et al., 2004; Reiser & Milne, 2014). Accurate and individualised feedback enables therapists to adopt effective strategies to enhance client outcomes (Ivers et al., 2012; Tracey et al., 2014). Research shows that feedback is most effective when it is distributed over a period of time on multiple occasions (Ivers et al., 2012). For example, three to four post-workshop feedback sessions prevents skill erosion among Motivational Interviewing trainees (Schwalbe et al., 2014). However, providing feedback using traditional methods is an expensive process for agencies and a time consuming job for supervisors. It can be even a more resource-intensive process when there are many therapists in a large scale training. New techniques, such as machine learning, are capable of quickly and cheaply analysing large-scale data, providing accurate individualised feedback.

Automated coding methods have been applied to large psychotherapy datasets up to 1553 sessions (Xiao et al., 2016). Once these models are trained, they can be repeatedly applied at very low cost (Xiao et al., 2016). They can reduce the likelihood of implicit bias of human decision-making (Lum, 2017), where the look or the sound of the therapist may contribute to errors in judgments. While some may doubt whether therapists would accept the feedback from machine learning models, preliminary feedback has been promising. Hirsch and colleagues (2018) provided ML based-feedback for 21 counsellors and trainees. The results of their qualitative study showed that counsellors were receptive to a computerised assessment, and were less defensive toward critical feedback from a machine than a human. It has also been documented that therapists are quite open to receiving machine learning feedback (Imel et al. 2019). In sum, machine learning models can cheaply provide objective feedback to therapists in a way that they are likely to find valuable.

Verbal Behaviour May Be a Good Candidate for Machine Learning. Interpersonal interactions in a therapy process involves a range of behaviours such as verbal behaviours (i.e., what is said) and non-verbal behaviours (such as prosody, body movements, biological changes). However, verbal behaviours are the primary channel of transferring information in dyadic interactions (Miller et al., 2003). Systematic reviews have shown that therapists' verbal behaviours are associated with various client outcomes, such as patient satisfaction and adherence to treatment (Golin et al., 2002; Howard et al., 2009). Most existing measures for assessing treatment fidelity focus on the words used by the therapist, rather than their tone or non-verbal behaviour (McGlinchey & Dobson, 2003; Miller et al., 2003; Moyers, Martin, et al., 2005). Verbal behaviour is also easy to code automatically, where even simple 'word-counting' methods can reliably and validly predict many psychological constructs (Pennebaker et al., 2003). Further, methods for automatic assessment of verbal behaviour are different from those for non-verbal or para-verbal (e.g., signal-processing features like tone, pitch, and pacing) behaviours. Many such tools have allowed for automated assessment of patient characteristics, such as diagnoses (Low et al., 2020). Emerging technologies may be able to code some non-verbal behaviour like sign language, but those technologies are not sufficiently advanced that they can code the nuanced non-verbal cues involved in psychosocial interventions. So, while non-verbal and para-verbal modalities are critical components of therapy, we focused on verbal interactions as an important and tractable machine learning task.

To analyze verbal behaviour, human coders are trained to identify specific therapy behaviours. The reliability of human-to-human codes are evaluated via a process called interrater reliability. Just as therapists drift, coders do too, where interrater reliability can decrease with fatigue or without frequent re-calibration (Atkins et al., 2012; Haerens et al., 2013a). Often when two humans code for fidelity using words therapists use, they are not perfectly aligned. Coders may overcome the 'coding drift' by meeting regularly to discuss their codes and instances of coder disagreement. However, human coding also faces other challenges such as being tedious, expensive and time consuming (Moyers et al., 2005). This means that human coding is an imperfect reference point, but a useful one to compare machine learning models against.

Proof-of-concept comes from many other fields in which machine learning has been found to reliably automate laborious tasks (Russell & Norvig, 2002). Ryan and colleagues (2019) have argued that machine learning is already good enough to assess the content and delivery of healthcare by doctors. They have been applied to predict language disorders (Can et al., 2012), and addiction and suicide behaviour (Adamou et al., 2018). In psychotherapy, they have been used to predict counselling fidelity (Atkins et al., 2014), empathy (Xiao et al., 2015a), and counsellor reflections (Can et al., 2016). A recent systematic review showed that 190 studies used machine learning methods to detect and diagnose mental disorders, indicating the applicability of machine learning in mental health research (Shatte et al., 2019). Similarly, Aafjes-van Doorn and colleagues (2020) did a scoping review of machine learning in the psychotherapy domain and showed that 51 studies applied machine learning models to classify or predict labeled treatment process or outcome data, or to identify clusters in the unlabeled patient or treatment data (Aafjes-van Doorn et al., 2021). Machine learning methods have also been used in psychiatry to parse disease models in complex, multifactorial disease states (e.g., mental disorders; Tai et al., 2019). When taken together, there are a number of domains in which machine learning models have been helpful in coding verbal behaviours, indicating they may be a powerful tool for psychotherapists and other helping professions.

Review Aims. The primary goal of this review is to assess how well machine learning performs as a method for assessing treatment fidelity using verbal aspects of therapist language. By conducting a systematic review, we were able to assess how well those models applied across studies and contexts. Models may only work well under a narrow set of conditions, and systematic reviews are able to assess those conditions more robustly than a narrative review. There are also some well-established best-practices that influence whether a machine learning model will generalise to new data (Luo et al., 2016). By assessing adherence to these guidelines, our review was able to indicate how well these models may generalise. Finally, we included all interpersonal interactions from helping professionals, even those outside psychotherapy (e.g., medicine, education), in order to assess whether machine learning models to assess communication and fidelity have been successfully implemented in nearby fields. In doing so, we could see whether models applied to medicine or education might be useful to consider in future psychological research. In sum, we sought to answer the following research questions:

1. Which automated coding methods have been used to analyse interpersonal verbal behaviours of helping professionals (with specific focus on fidelity in psychotherapy)?
2. How accurate are machine learning methods?
3. To what extent have studies applying automated coding methods adhered to best-practice guidelines for machine learning?

Methods

We report this systematic review in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement (Moher et al., 2009).

Protocol and Registration

We prospectively registered the protocol in the Prospective Register of Systematic Reviews (PROSPERO registration number: CRD42019119883).

Eligibility Criteria

In this review, we included studies meeting the following criteria:

1. The participants or population studied were helping professionals. A helping professional engages in “a professional interaction with a client, started to nurture the

growth of, or address the problems of, a person's physical, psychological, intellectual, or emotional constitution" (Graf et al., 2014, p. 1). Examples of helping professionals are psychotherapists, counsellors, doctors, nurses, teachers, and social workers.

2. They measured verbal interpersonal interactions between helping professionals and clients (e.g., clinician and client, or teacher and student).
3. They analysed the helping professionals' verbal behaviour (i.e., language) that occurred during interpersonal interactions.
4. They used an automated method for coding behaviour. Coding refers to the process of either rating or categorising an interpersonal interaction on at least one variable. Automated coding methods refer to the methods which code the input data, without manual interference in the coding process. The input data for such systems could be transcripts, audio tracks, or video clips (with audio included). Codes are labels that are used to represent certain behaviours, and they may vary in their level of granularity or specificity and concreteness (ranging from physically to socially based codes; Bakeman & Quera, 2011).
5. Both peer reviewed and grey-literature (e.g., conference papers, theses) were eligible for inclusion.
6. Papers written in any language with title and abstract in English were included.
7. Any design, location or year were included.

Exclusion Criteria

We excluded studies if:

1. Participants were not helping professionals.
2. They analysed interprofessional interactions (e.g., doctors interacting with nurses).
3. They analysed interpersonal interactions using only aspects other than language (i.e., facial expressions, body posture and gestures).

4. They used semi-automated methods (where the final results still required some human coding) or manual methods (where a human is needed to code the behaviour).
5. They were published abstracts, without a full-length paper.

Search Strategy and Information Sources

To develop the search strategy, we created an initial pool of target papers that met the inclusion criteria. We conducted forward and backward citation searching on this initial pool (Hinde & Spackman, 2015) to identify six more papers meeting the eligibility criteria. We extracted potential search terms from these 11 papers by identifying key words from the title and abstract (Hausner et al., 2016). The final search strategy involved keywords and their MeSH terms or synonyms from four main groups including 'participants' (e.g., teacher or doctor), 'measurement' (such as assessment or coding), 'automated coding method' (e.g., Natural Language Processing or text mining), and 'type of behaviour' (e.g., fidelity or interaction). The search did not have any exclusion terms (see Appendix A.2 for full search details and included papers).

We performed the search within PubMed, Scopus, PsycINFO, Education Source, ERIC, CINAHL Complete, Embase, SPORTDiscus, and Computers and Applied Sciences Complete databases. We performed the last search on the 21st of February 2021. To test the sensitivity of our strategy, we first confirmed that the identified records included 11 target papers described earlier. We then searched the first 200 results on Google Scholar to identify potentially relevant studies not indexed in electronic databases.

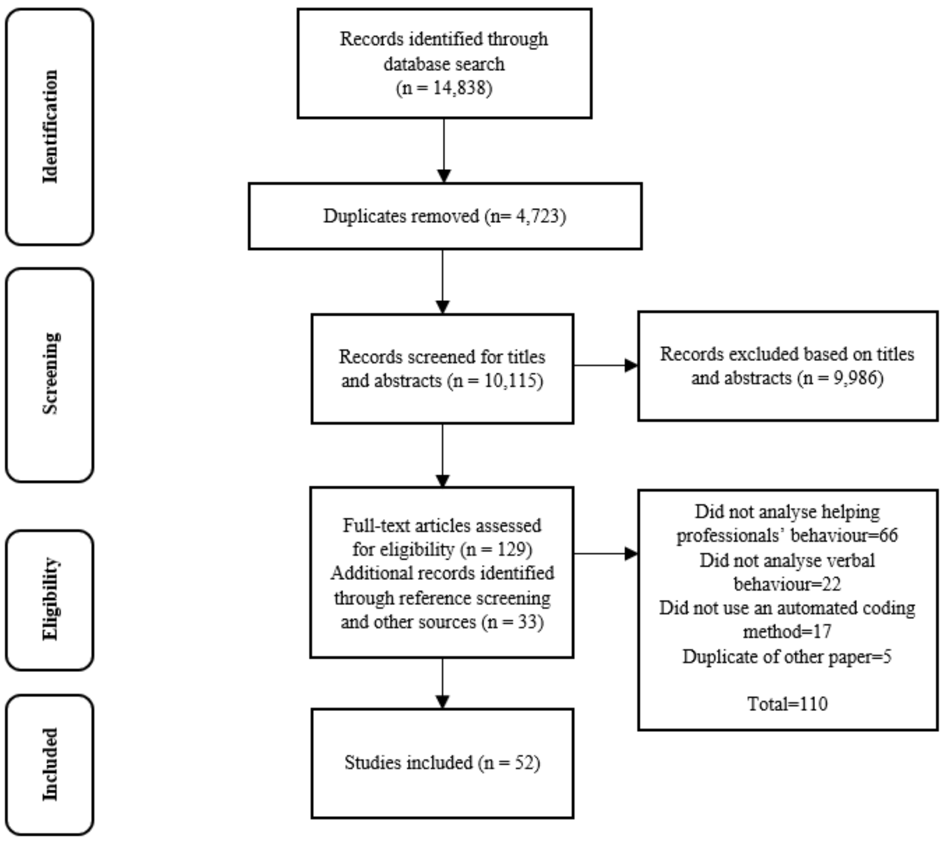
We conducted forward and backward citation searching on studies that passed fulltext to identify related papers which did not appear in the systematic search (Greenhalgh & Peacock, 2005; Hinde & Spackman, 2015). We also emailed the first author of included papers and known experts in the automated coding of verbal behaviour to identify any unpublished manuscripts.

Study Selection

We imported search results into Covidence software (Babineau, 2014). We dealt with studies in two steps. First, we screened the titles and abstracts of the studies according to the pre-defined inclusion criteria. If the title or abstract did not provide enough information to decide, we moved the record to full-text screening. Second, we reviewed full texts of articles for final inclusion. At each stage, two reviewers (AA and MS, or AA and DA) independently made recommendations for inclusion or exclusion. We resolved any discrepancies in study selection at a meeting. Then, we resolved any conflicts by consulting with a third reviewer (MN). The PRISMA flow diagram (Figure 2.1) provides detailed information regarding the selection process.

Figure 2.1

PRISMA Flow Diagram of the Study Selection Process



Data Collection Process

We developed a data extraction form for this review to focus on the applied automated coding methods and their performance. We first tested the form by extracting data from four randomly selected papers. Two researchers (AA and MS or AA and DA) then independently extracted data from each study and organised it into tables to display themes within and across the included studies. Any discrepancies from the data extraction were discussed between the reviewers. In the case of unresolved disagreements, a third reviewer (MN) was consulted.

Adherence to Best-Practices in Machine Learning

We assessed study quality using a tool based on the "Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View" (Luo et al., 2016). This tool was used to judge the extent to which studies adhered to best-practice guidelines. The original checklist contained 51 items and investigated the quality of papers based on the information in each section of a paper. The checklist was used in two ways. One researcher (AA) assessed all 51 items. We refined this checklist by identifying the core items related to performance of automated coding methods. Of the 51 items, nine were related to the performance (see identified items in Table 2.4, and the complete checklist in the online materials available at https://osf.io/9juhd/, Supplementary file 3); the others related to the reporting in the manuscript (e.g., three items are whether the abstract contains background, objectives, and data sources sections). The other researcher (MS/DA) assessed the core checklist. Specifically, the two researchers independently assigned the label "Yes" if the requisite information was described explicitly and "No" if the information was not adequately described. Rather than reporting a summary score (e.g., "high" or "low quality"), we followed Cochrane guidelines that recommend reporting quality scores for each item of the quality assessment checklist (Macaskill et al., 2010).

Results

Study Selection and Results of Individual Studies

Our systematic search resulted in 14,838 records. We removed 4,723 duplicates, with 9,986 papers remaining for title and abstract screening. Thirty-three further records were added by other methods (e.g., forward and backward searching). Fifty-two papers met the inclusion criteria and were included in this review (see Figure 2.1). All the included papers were written in English. Online materials (Supplementary file 4) available at https://osf.io/9juhd/ summarises the information from individual studies.

Synthesis of Results

Most of the studies were conducted in psychotherapy settings (k=34, 65.3%) and involved counsellors, psychologists, or psychiatrists. Nine studies were conducted in a medical care setting (16.6%) and included physicians or nurses. Ten studies (18.5%) were conducted in education contexts and involved school teachers. Of the 53 studies, 23 (41.5%) examined Motivational Interviewing (Miller & Rollnick, 1991) with the rest of the studies scattered across different modalities (one paper included two studies, for details see Table 2.1).

Table 2.1

Context of Study

Context	Psychotherapy	Medical care	Education
Studies	Counselling, Motivational Interviewing (Counsellors), (Atkins et al., 2014; Can et al., 2012; Can et al., 2015; Can et al., 2016; Carcone et al., 2019, Study 1; Chakravarthula et al., 2015; Gibson et al., 2016; Gibson et al., 2017; Gupta et al., 2014; Hasan et al., 2019; Hasan et al., 2018; Imel et al., 2015; Perez-Rosas et al., 2017; Perez-Rosas et al., 2019; Singla et al., 2018; Tanana et al., 2016; Xiao et al., 2012; Xiao et al., 2015; Xiao et al., 2016; Xiao, Can, et al., 2016; Chen et al., 2019; Gibson et al., 2019; Cao et al., 2020)	Medical care, provider-patient clinical interactions (Carcone et al., 2019, Study 2; park et al., 2019)	Education (teachers) (Blanchard et al., 2016; Blanchard et al., 2016; Donelly et al., 2016; Donnely et al., 2016; Donnelly et al., 2017; Samei et al., 2014; Samei et al., 2015; Wang et al., 2014; Song et al., 2020; Suresh et al., 2019)
	Counselling, (Counsellors), (Althoff et al., 2016; Gallo et al., 2015; Gaut et al., 2017; Malandrakis et al., 2015; Mieskes et al., 2018; Nitti et al., 2010; Salvatore et al., 2012; Velasquez et al., 2018; Goldberg et al., 2020; Flemotomos et al., 2018)	Medical care, (nurses) (Lacson et al., 2005)	
	Counselling (Psychiatrists), Howes et al., 2013	Medical care (physicians, nurses, physician assistants) (Mayfield et al., 2014)	
		Medical care (Oncologists) (Sen et al., 2018)	
		Medical care (physicians) (Angus et al., 2012; Wallace et al., 2013; 2014; Park et al., 2021)	
Total*	35 (64.8%)	9 (16.6%)	10 (18.5%)

Note. *One study was performed in two different contexts

Note. *One study was performed in two different contexts

Predicted Outcomes

Studies in the psychotherapy context aimed to predict the fidelity to a prescribed therapeutic process (k=28, 82.3% of psychotherapeutic studies). In medical care settings, the aim was to identify clients' symptoms (k=1), topics discussed in conversations (k=5), or conversational patterns (k=5). In educational contexts, studies aimed to predict the number of teacher questions (k=5) and the type of classroom activities (e.g., discussion, lecture, or group work, k=5).

Behavioural Coding Measures and Automated Coding Methods

Many studies used automated coding to implement pre-existing behavioural coding measures. Behavioural coding measures were usually designed to measure adherence to the practice guidelines or instructions. The majority of studies used a behavioural coding measure (for details, see Table 2.2). The most frequently applied coding measure was Motivational Interviewing Skills Code (k=11, Miller et al., 2003), followed by the Motivational Interviewing Treatment Integrity measure (k=7, Moyers, Martin, et al., 2005). Seven studies used a coding system to code whether teachers asked questions, provided instructions, or facilitated small-group activities (Nystrand et al., 2003).

Table 2.2

Frequency of Behavioural Coding Measures Used in Included Studies

Behavioural Coding Measure	Frequency
Motivational Interviewing Skill Code	14
Motivational Interviewing Treatment Integrity	7
Nystrand and colleagues coding scheme (2003)	7
Minority Youth-Sequential Code for Observing Process Exchanges	3
Generalized Medical Interaction Analysis System	3
Diagnostic and Statistical Manual of Mental Disorders - 4th edition	2
A coding manual developed in a previous study (in Prado et al., 2006; Stigler et al., 2000)	2
Cognitive Therapy Rating System (CTRS)	2
Cognitive therapy scale for psychosis (in Lecomte, Kingdon, and Munro-Clark, 2017)	1
Accountable Talk framework (Michaels, O’Connor, and Resnick, 2008).	1

Multi-Dimensional Interaction Analysis coding system	1
Did not apply a previously established behavioural coding system	12

Note: Some studies used more than one behavioural coding measure.

In this context, the machine learning methods were designed to automatically assign codes from the behavioural coding measures to overt interactions recorded in the dataset (e.g., words/utterances). Most studies assessed more than one machine learning method; the most frequently applied were Support Vector Machine (k=8), Random Forests (k=7), Logistic Regression (k=7), J48 classifiers (a type of decision tree, k=6), Maximum Entropy Markov models (k=5), and Naive Bayes (k=5; for details, see Table 2.3).

Table 2.3

Automated Coding Methods

Automated Coding method	Frequency*	Citations
Support Vector Machine	8	Carcone et al., 2019 (Study 1 and 2); Howes et al., 2013; Perez-Rosas et al., 2017; Perez-Rosas et al., 2018; Xiao et al., 2015; Park et al., 2019; Flemotomos et al., 2018
Random Forest	7	Carcone et al., 2019; Imel et al., 2015; Mieskes et al., 2018; Blanchard et al., 2016(a); Blanchard et al., 2016(b); Donnelly et al., 2017; Wang et al., 2014
Logistic Regression	7	Park et al., 2019; Sen et al., 2018; Donnelly et al., 2017; Blanchard et al., 2016 (a); Blanchard et al., 2016 (b); Park et al., 2021; Mayfield et al., 2014
J48 (Decision Tree)	6	Carcone et al., 2019; Howes et al., 2013; Blanchard et al., 2016(a); Blanchard et al., 2016(b); Donnelly et al., 2017; Samei et al., 2014;
Maximum Entropy Markov	5	Can et al., 2012; Can et al., 2016; Gupta et al., 2014; Xiao, Can, et al., 2016; Xiao., Huang et al., 2016;
Naive Bayes	5	Carcone et al., 2019; Blanchard et al.,

		2016(a); Donelly et al., 2016 (a); Donnely et al., 2016 (b); Donnelly et al., 2017;
Recurrent Neural Networks	5	Hasan et al., 2018; Singla et al., 2018; Blanchard et al., 2016 (a); Park et al., 2021; Gibson et al., 2017;
Hidden Markov Model	4	Althoff et al., 2016; Can et al., 2012; Hasan et al., 2019; Hasan et al., 2018;
K-Nearest Neighbours	4	Blanchard et al., 2016(a); Sen et al., 2018; Blanchard et al., 2016(b); Donnelly et al., 2017;
Conditional Random Field	4	Can et al., 2015; Carcone et al., 2019; Wallace et al., 2014; Park et al., 2019
Bi-directional Long Short Term Memory (Bi-LSTM)	3	Chen et al., 2019; Gibson et al., 2019; Suresh et al., 2019;
Labelled Topic Model	2	Atkins et al., 2014; Imel et al., 2015
Bayesian Network	2	Blanchard et al., 2016 (a); Blanchard et al., 2016(b);
Gared Recurrent Unit (GRU)	2	Cao et al., 2020; Park et al., 2019;
30 models were used once each.**	1 each (30 in total)	Angus et al., 2012; Carcone et al., 2019; Chakravarthula et al., 2015; Gallo et al., 2015; Gaut et al., 2017; Gibson et al., 2016; Gibson et al., 2017; Hasan et al., 2018; Howes et al., 2013; Imel et al., 2015; Lacson et al., 2005; Malandrakis et al., 2015; Nitti et al., 2010; Salvatore et al., 2012; Tanana et al., 2016; Velasquez et al., 2018; Wallace et al., 2013; Xiao et al., 2012; Xiao., Huang et al., 2016; Xiao, Can, et al., 2016; Samei et al., 2015; Park et al., 2019; Song et al., 2020; Goldberg et al., 2020;
Unique Models= 41	All the used models=76	

Note. *Some studies applied more than one coding method. We reported all the specific models that were applied in the studies. Some models might be variations of another model.

** The models were: Activation-based Dynamic Behaviour Model (ADBM) using Hidden Markov Model, AdaBoost, Automated Co-occurrence Analysis for Semantic Mapping (ACASM), Boostexter tool, Deep Neural Networks, DiscLDA, Discourse Flow Analysis (DFA), Discrete Sentence Features using Multinomial Logistic Regression, Discursis software, Fidelity Automatic RatEr (FARE system), Joint Additive Sequential (JAS) model using Log-linear classifier, Labeled Latent Dirichlet Allocation, Lasso Logistic Regression (LLR), Latent

Dirichlet Allocation, Likelihood-based Dynamic Behaviour Model (LDBM) using Hidden Markov Model, Linear Regression, Markov Chain, Markov-Multinomial, Maximum Likelihood Classifier with Universal Background Model (UBM) and Kneser-Ney algorithm, Maximum Likelihood Model with Kneser-Ney algorithm, Naive Bayes-Multinomial, RapidMiner, Recurrent Neural Networks with Gated Recurrent Unit (GRU), Recursive Neural Network (RNN), Ridge Regression model, Static Behaviour Model (SBM) using Universal Background Model, Hidden Markov Model Logistic Regression (HMM-LR), Hidden Markov Model-Support Vector Machine (HMM-SVM), Hidden Markov Model-Gated Recurrent Unit (HMM-GRU), Convolutional Neural Network - Bidirectional Long Short Term Memory (CNN-BiLSTM) model.

Which Methods Performed Best?

In Appendix A.3, we report the predictive performance of each method (e.g., F1-score measure for the Support Vector Machine in Xiao et al., 2015a is 0.89). We also reported a brief description of each coding method and accuracy measures in the Appendix A.1. Methods generally performed well in terms of their agreement with human coders. Overall, Kappa ranged from 0.24 to 0.66, with all but one study (Samei et al., 2014) falling between 0.38 and 0.66. These results suggested fair to excellent levels of agreement, compared with established thresholds for Kappa used for human-to-human agreement (Landis & Koch, 1977). Accuracy—meaning the ratio of correctly predicted codes to the total number of predictions—was greater than 50% in all studies and sometimes higher than 80% (e.g., Chakravarthula et al., 2015; Wang et al., 2014; Xiao et al., 2016).

Support Vector Machine methods generally performed well. For example, Xiao and colleagues (2015) found that the Support Vector Machines methods performed almost as well as trained coders. Similar results were reported in other studies (e.g., Flemotomos et al., 2018; Pérez-Rosas et al., 2017, 2019). Most studies only examined one type of method's performance. In one study that directly compared different methods on the same dataset, Support Vector Machines outperformed seven alternative method strategies in terms of agreement with human coders and accuracy (Carcone et al., 2019).

Because few studies examined the performance of methods when transferred to other similar settings—for example, with similar predictors and outcomes but different participants—we are unable to ascertain whether any particular method predicted new data better than others. There were three studies that compared the performance of methods but did not report the predictive performance of all the tested methods and only chose the best performing method (Blanchard et al., 2016a, 2016b; Donnelly et al., 2017). Only one study developed a Support Vector Machine method in psychotherapy and applied it on new data from another context (i.e., medicine; Carcone et al., 2019). The method performed well, achieving a substantial level of agreement with human coding.

Larger Datasets Lead to More Accurate Performance

Dataset sizes ranged from 13 sessions (Wang et al., 2014) to 1,235 sessions (Goldberg et al., 2020). When the dataset size was larger, methods performed more accurately. For example, Imel and colleagues (2015) analysed more than 9 million words and the method achieved an accuracy of 87% (using a Random Forest). Similar results were reported in other studies with large datasets (e.g., Gaut et al., 2017; Xiao et al., 2015a, 2016). Perez-Rosas and colleagues (2019) showed that as they increased the amount of data in their training set they observed significant improvement in prediction accuracy. Aligned with this finding, frequently observed codes (i.e., categories) in a dataset were predicted more accurately, while low base rate codes were predicted less accurately (e.g., Can et al., 2015; Cao et al., 2019; Carcone et al., 2019; Gibson et al., 2017; Tanana et al., 2016; Wallace et al., 2014). An example of frequently observed code is 'open questions' and an example for low base rate codes is 'confrontational statements'.

The Fewer the Codes the More Accurate the Performance

Methods classified data into codes, with the number of codes ranging from two (Blanchard et al., 2016b; Xiao et al., 2015a) to 89 (Gaut et al., 2017). When the number of codes decreased, performance of the method increased, and vice versa. Carcone and colleagues (2019) showed that the methods performed better in 17-code prediction than 20-code prediction, and 20-code prediction was superior to 41-code prediction. Similar results were reported in other studies that directly compared coding frameworks of differing complexity (e.g., Gallo et al., 2015; Gibson et al., 2016). When methods were simplest (i.e., two codes), accuracy was greater than 80% (e.g., Blanchard et al., 2016b; Chakravarthula et al., 2015; Gallo et al., 2015; Pérez-Rosas et al., 2019; Xiao et al., 2016). When the number of codes was higher, prediction was less accurate (i.e., accuracy=54% with 41 codes in Carcone et al., 2019; accuracy=66% with 20 codes in Howes et al., 2013).

More Concrete and Less Abstract Codes Lead to Better Performance

The conceptual meaning of the codes affects the predictive performance of methods. Methods accurately predicted some types of codes. For example, questions (e.g., a counsellor or teacher asking questions to gather information, such as “How do you feel about that?”) and facilitation (i.e., simple utterances that function as acknowledgements and a cue to continue speaking, such as “hmm-mm”) seem to be conceptually concrete. These codes were predicted more accurately than conceptual abstract codes, such as empathy (Atkins et al., 2014), confrontation, and advising (Imel et al., 2015; Tanana et al., 2016).

Session-Level Prediction Is More Accurate Than Utterance-Level Prediction

Utterance-level prediction refers to the prediction of a small unit of spoken words that have a specific meaning (i.e., complete thoughts). For instance, “you feel overwhelmed” is an utterance that may signal reflective listening. Session-level prediction refers to the prediction of a behaviour or skill over a session. For example, in Motivational Interviewing Treatment Integrity coding measure, the empathic quality of the provider is rated on a 1-5 Likert scale, taking the entire session or segment of the session into account (Moyers et al., 2016). Session-level prediction may also code whether the therapist implemented a specific behaviour (e.g., reflective listening) frequently (e.g., 10/10) or rarely (0/10). Compared with utterance-level prediction, Tanana and colleagues (2016) showed that the session-level prediction results had stronger concordance with human-based coding. Atkins and colleagues

(2014), and Park and colleagues (2019) reported similar results, where the session-level prediction was generally closer to human coding rather than utterance-level prediction.

Quality of Reporting Within Studies

Results of our study quality assessment can be found in Table 2.4. Inter-rater reliability analysis of the quality assessment among this systematic review team showed agreement on 89% of the instances assessed by two independent reviewers. We resolved discrepancies by discussion between the two researchers (AA and MS or AA and DA) and consultation with a third reviewer (MN).

Table 2.4

Quality Assessment

Item/Study	Clarified the clinical setting for the target predictive model?	Described the modelling context in terms of facility type, size, etc.	Defined a measurement for the prediction goal	Defined the success criteria for prediction	Defined the observational units	Described the data pre-processing	Define model validation strategies	Reported the predictive performance of the final model	Reported variables to be predictive of the response variable
Althoff et al., 2016	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes
Angus et al., 2012	Yes	Yes	Yes	No	Yes	Yes	No	No	No
Atkins et al., 2014	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No
Blanchard et al., 2016a	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No
Blanchard et al., 2016b	Yes	Yes	No	No	Yes	No	Yes	Yes	Yes

Can et al., 2012	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No
Can et al., 2015	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No
Can et al., 2016	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No
Carcone et al., 2019	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Chakravarthula et al., 2015	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No
Donnelly et al., 2016a	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes
Donelly et al., 2016b	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No
Donnelly et al., 2017	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes
Gallo et al., 2015	Yes	Yes	Yes	Yes	Yes	No	No	Yes	No
Gaut et al., 2017	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Gibson et al., 2016	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No
Gibson et al., 2017	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No
Gupta et al., 2014	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No

Hasan et al., 2018	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No
Hasan et al., 2019	Yes	Yes	Yes	No	Yes	Yes	No	No	Yes
Howes et al., 2013	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Yes
Imel et al., 2015	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No
Lacson et al., 2005	Yes	Yes	No	Yes	Yes	No	No	Yes	No
Malandrakis et al., 2015	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Mayfield et al., 2014	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No
Mieskes et al., 2018	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Nitti et al., 2010	Yes	Yes	No	No	Yes	Yes	No	No	Yes
Perez-Rosas et al., 2017	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Perez-Rosas et al., 2019	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Salvatore et al., 2012	Yes	Yes	No	No	Yes	Yes	Yes	Yes	No

Samei et al., 2014	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes
Samei et al., 2015	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes
Sen et al., 2018	Yes	Yes	No	No	Yes	No	Yes	Yes	No
Singla et al., 2018	Yes	Yes	Yes	No	Yes	Yes	No	Yes	No
Tanana et al., 2016	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No
Velasquez et al., 2018	Yes	Yes	No	No	Yes	No	No	No	No
Wallace et al., 2013	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Wallace et al., 2014	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Wang et al., 2014	Yes	Yes	No	No	Yes	No	Yes	Yes	Yes
Xiao et al., 2012	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Xiao et al., 2015	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Xiao, Can, Gibson et al., 2016	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes

Xiao, Huang, Imel et al., 2016	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No
Chen et al., 2019	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No
Gibson et al., 2019	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No
Park et al., 2019	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Flemotomos et al., 2018	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Cao et al., 2020	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Park et al., 2021	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes
Song et al., 2020	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Suresh et al., 2019	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Goldberg et al., 2020	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	No
Sum of ‘Yes’ items	52	52	43	18	52	28	44	48	19

We report quality assessment results for each item of the core checklist (nine items). All the papers reported the clinical setting, dataset details, and observational units. Forty-five papers (86.5% of studies) coded behaviours using a behavioural coding measure. These types of concrete guidelines facilitate utterance level comparison. Twenty-eight papers reported data pre-processing (53.8% of the studies), which improves performance of a method by removing outliers or poor quality data (e.g., removing very low quality voice recordings; García et al., 2014). Thirty-four papers (64.1% of studies) validated the methods using some form of cross-validation (where a method is trained on a dataset and tested on a unseen set of observations; Browne, 2000). Yet, only Eleven papers (21.1% of studies) applied a hold-out 'train and test' method. Studies that do not test the accuracy on unseen data can overfit the data to the training set, and give misleading estimates of how accurately the method can predict new data (Yarkoni & Westfall, 2017). Eighteen papers (34.6% of studies) reported success criteria (e.g., mean-squared error), which help to interpret performance of a method. Relative importance of predictor variables (e.g., which feature is most important in predicting the outcome variable) were reported in 19 papers (36.5% of studies). For full details about each quality indicator, see Table 2.4.

Discussion

Our systematic review found that several automated coding methods have been applied to assess fidelity in psychological interventions. We also identified many methods used to analyse verbal interactions of other helping professionals, not just therapists. These methods generally demonstrated promising results with accuracy comparable to human coding. Methods performed better on large datasets, coding frameworks with fewer behaviours, and verbal behaviours that represent concrete (rather than abstract) codes. However, studies rarely reported adherence to best-practice machine learning guidelines, meaning that the machine learning models may not generalise well to new interactions with new clients, reflecting a deficit in the field.

Methods showed promising performance in automatic annotation of therapist' verbal behaviour, including treatment fidelity to a number of models (most frequently Motivational Interviewing). This result suggests machine learning could reduce financial costs of traditional methods. Doing so would improve the scalability and efficiency of behavioural coding for assessment and feedback on treatment fidelity. When directly compared with other methods, the Support Vector Machines method showed superior performance and appeared to be an appropriate method for generalisability purposes (Carcone et al., 2019). The higher performance of the Support Vector Machines method was also reported in other studies in the similar applications (Hasan et al., 2016; Kotov et al., 2014). This method might have potential in less-explored contexts such as fidelity for cognitive behaviour therapy or acceptance and commitment therapy, because the machine learning models efficiently process sparse, high-dimensional data and non-linearities with few interactions.

Having said that, the field of machine learning is advancing quickly and the methods reported here may not reflect the current state-of-the-art. For example, Kaggle's machine learning competitions have recently been dominated by Extreme Gradient Boosting or Neural Network methods (Abou Omar, 2018). New, powerful, natural language models contain up to 175 billion parameters and require only a few pieces of training data (Brown et al., 2020). We expect that automated coding methods will become even more powerful, and better-able to manage ambiguity, once researchers start implementing these cutting-edge methods. Our findings were restricted by the small number of studies that directly compared different machine learning methods; therefore, caution should be taken when generalising the predictive performance of these methods to other cases. Researchers in this area could help accelerate the field by transparently reporting which models were tested and discarded, and why. It is common practice in machine learning to test a number of models using cross-validation on the training set (Cawley & Talbot, 2010); we were therefore surprised to see so few head-to-head comparisons reported. It is possible that researchers only reported the performance of a model that performed best with their data. This is concerning because few studies reported how well the models predicted unseen data on a hold-out, ‘test set’ and thus the risk of over-fitting was potentially high.

There were rare cases where automated coding methods did not perform well (Gallo et al., 2015; Samei et al., 2014). While the method itself can be an important factor in prediction accuracy, there are important conditional factors, such as dataset size, that affect a method’s accuracy (Yarkoni & Westfall, 2017). Considering these conditions, it was not easy to provide a fair comparison between statistical models because the choice of model was often confounded by differences in samples and prediction objectives. In the following section, we present a cautious overview of the factors that influence the methods’ predictive performance and provide suggestions for future research and practice.

While determining the appropriate size of a dataset remains a matter of debate, large datasets support training, testing, and generalization of predictions in new datasets (Yarkoni & Westfall, 2017). Future studies could identify whether or not more data are needed by looking at the learning curves, which show whether the method can be adequately fit with the data available (Perlich, 2009). In general, our results showed that larger datasets lead to better performance. This finding is in line with previous studies where machine learning algorithms generally performed better on larger datasets (Domingos, 2012). It is important to note, however, that additional data have diminishing returns. As such, it is important for analysts to monitor method performance as sample sizes increase in order to maintain reasonable cost-benefit ratios (Ng, 2019).

Another factor influencing methods’ performance is the number of codes a method is built to predict. Methods generally performed worse when the number of codes increased (e.g., Gallo et al., 2015; Hasan et al., 2016). As such, we recommend analysts carefully consider which codes are most critical as a means of increasing method performance. When learning curves indicate that data is under-fit, then authors could consider using fewer codes (e.g., by collapsing conceptually similar codes) to allow for more reliable methods.

Codes with simple conceptual meaning were predicted more accurately (e.g., open-ended questions), while complicated codes were predicted weakly (e.g., informing with permission from the client vs. informing without permission). Researchers might consider the trade off between the lower prediction accuracy for complicated codes and the higher costs of coding them using alternative methods (e.g., manual coding). Similarly, codes that can be objectively identified in a transcript (e.g., questions, affirmations, and facilitations) are likely to be more easily coded than those that require inference and subject-matter expertise.

Many accurate methods in this review were applied in the Motivational Interviewing context. The behavioural coding systems for Motivational Interviewing are well defined and more reliably coded than many other therapeutic approaches (Miller & Rollnick, 1991). This may be because Motivational Interviewing explicitly prescribes a number of conversational devices (e.g., reflections, affirmations, open questions) to be used in session, where other practices are less prescriptive regarding the conversation process and more focused on the content of discussion (e.g., the client's idiosyncratic negative automatic thoughts). Similarly, the techniques prescribed by motivational interviewing may occur hundreds of times a session (e.g., reflective listening). Core techniques from other treatment approaches may only happen once per session (e.g., checking homework). As a result, machine learning methods may be less reliable where behavioural codes are less clear, like in other psychological treatment approaches (e.g., cognitive-behaviour therapy).

Finally, methods tend to perform poorly when codes are constructed at the utterance-level; the overall prediction of a code was more reliable over a session. Part of the reason for this arises from the difficulty of utterance-level coding tasks—even for human coders—if they do not rely on the prior or subsequent utterances (Tanana et al., 2016). Without context, it is difficult to know whether “your drinking is a problem” is an empathic response to the client’s self-awareness or a controlling, unsolicited prescription. As a result, it is more reasonable to rely on the overall prediction results over a session rather than each individual utterance. Recently, Cao and colleagues (2019) investigated the prediction of therapist utterance labels by taking the context of the utterance into consideration. They found that by increasing the history window size (i.e., by accounting for the last 8 utterances), categorization accuracy improved (Cao et al., 2019). This indicates that providing machine learning with more context may improve the accuracy of models. The other reason for poor performance at utterance-level prediction compared to session-level prediction may be that, across a session, the machine-learning task is closer to a regression problem than a classification problem. That is, it may be hard to classify a moment as ‘empathic’ from a set of words, but it may be easier to correlate ratings of empathy with the frequency of specific words across an entire session (e.g., “you feel…”, “it sounds like…”).

Atkins and colleagues (2014) presented the potential factors impacting the accuracy of Topic Models in predicting client and therapist codes in the Motivational Interviewing Skill Code. Like our review, they argued that models worked less accurately at utterance (i.e., talk-turn) level than at session level. They also stated that more abstract codes were weakly predicted than more concrete ones. However, their findings only focused on one of the many psychosocial interventions (motivational interviewing), and our systematic review identified other factors which are likely to influence the performance of machine learning methods. Particularly, this systematic review showed that larger datasets and more frequently observed codes lead to better prediction accuracy. Also, fewer target behaviours leads to higher accuracy. Further, other factors impact the predictive power of a model, such as the ML model selection process, pre-processing, and validation method.

Potential Applications

Specific and immediate feedback is essential to the development of skills across domains (Kahneman & Klein, 2009). Feedback works best when it is provided several times, spaced over a period of time (Ivers et al., 2012). However, providing individualised, distributed, and prompt feedback multiple times for a big group of therapists can be prohibitively expensive. Automated coding methods showed promising results in analysing helping professionals' language, so they can be used to provide feedback and improve practitioners' skills. Our systematic review shows that automated coding methods provided accurate estimation of treatment fidelity, including all three components (adherence, differentiation, and competence; Rodriguez-Quintana & Lewis, 2018). In motivational interviewing, for example, automated methods were able to code adherence to therapeutic strategies (e.g., affirming change), differentiation of proscribed strategies (e.g., use of closed questions; Tanana et al., 2016), and competence in delivery (e.g., session-level empathy ratings; Gibson et al., 2016). Specific, prompt feedback on all three of these may be useful for therapists. In the medical care setting, automated coding methods identified conversation patterns and discussed symptoms. In the education context, automated coding methods successfully predicted the number of questions teachers asked and the types of class activity they set. These automated methods are well tolerated (Skipp & Tanner, 2015). Imel and colleagues (2019) used automated coding methods to provide prompt feedback on therapists' performance in a laboratory setting. Therapists found the provided feedback representative of their performance and easy to be understood. Psychologists were shown to be more receptive to computerised feedback than from a supervisor (Hirsch et al., 2018; Imel et al., 2019). We are aware of only a few commercially available tools for assessing the fidelity of psychosocial interventions. For example, Atkins and colleagues deployed models (Imel et al., 2015; Tanana et al., 2016; Xiao et al., 2015b) for automatic coding of therapy sessions including CBT and motivational interviewing (Tanana, 2021). However, the dearth of publicly available tools reveals an opportunity for better collaboration between research and industry and improved knowledge translation.

From a research perspective, machine learning may allow for more affordable, reliable, scalable assessments of treatment fidelity. There is a substantial outlay in the initial annotation of therapy transcripts, but once this annotation is complete for a large trial, the data can be easily used to assess fidelity in other trials. The heterogeneity in fidelity assessment tools does add another level of difficulty for many modalities, like cognitive behavioural therapy, acceptance and commitment therapy, or interpersonal psychotherapy. If studies continue to use different assessments of treatment fidelity, then the generalisability of the machine learning models will be small. If the research community for each of these therapies agreed upon a set of core principles of change that were observable in therapy, then more annotated data would be available to train automated fidelity assessments for these therapies. In health, a number of delphi studies have been conducted that allowed experts to reach consensus on both a-theoretical and theory-driven strategies (Michie et al., 2013; Teixeira et al., 2020). Using these taxonomies, or more consistent use of a smaller number of fidelity assessment (e.g., Motivational Interviewing Skill Code; Miller et al., 2003; Motivational Interviewing Treatment Integrity; Moyers, Martin, et al., 2005), does lay the platform for machine learning methods of automated coding.

This research, however, needs to be careful to build models that perform well on future data, not just the data included in the original study. Assessing model fit on new data is a primary difference between predictive methods (i.e., machine learning) and more traditional explanatory modelling in research contexts (Breiman, 2001). Decision-rules that work in one dataset may not work with future data. For example, Google Flu Trends was able to predict historial flu rates from their search data, but it failed to accurately predict future data because methods became too sensitive to noise in the historical data (Lazer et al., 2014). To avoid these traps, machine learning experts identified a set of best-practice guidelines (Luo et al., 2016), which we used to evaluate studies. Our review found that few studies met these criteria. For example, guidelines recommend using a section of available data to refine the method (e.g., 70% of participants), but new data (e.g., 30% of participants), not used to refine the method, should be used for testing the final method (Luo et al., 2016; Yarkoni & Westfall, 2017). Only 21.1% of studies tested their methods on hold-out data. This is despite testing methods on novel data being an essential measure of method performance in machine learning. Six studies (11.1%) did not report how they refined their method at all (i.e., the validation process). Without transparently reporting these processes, readers cannot assume that machine learning methods will work on future data. Similarly, 46.2% of studies did not report if or how they undertook pre-processing of data. Pre-processing involves the cleaning and rescaling of data which usually occurs before training the method (García et al., 2014). Without these details, methods are not reproducible. While the general conditions of the studies were reported (e.g., where authors got the data and how much data they had), future predictive methods will be more useful, accurate, and generalisable if studies adhere to best-practice guidelines.

Limitations

The studies in this review used a wide variety of accuracy measures, behavioural coding measures, and outcomes which made it difficult to compare the methods. We could have calculated a common metric with a confusion matrix. Confusion matrices represent the predictive results of each code in utterance level (i.e., how many utterances predicted correctly or incorrectly), but only nine studies (three studies in psychotherapy and six studies in education) reported such a matrix. Another limitation was that treatment is a collaborative dialogue, but we only analysed the helping professionals' language. Some studies analysed both helping professionals' and clients' language, and methods that predicted both may be useful for clinicians and researchers to assess fidelity (e.g., did the technique produce the desired outcome). Also, predictive performance of a method might be different when analysing the clients' language, so future reviews could assess the methods used to automatically annotate client/patient language. Similarly, we excluded studies that only focused on signal-processing models of para-verbal behaviour, or object-classification models of non-verbal behaviour from video. Both non-verbal and para-verbal behaviour are important components of therapy, particularly with respect to common factors like therapeutic alliance. Future reviews may want to assess whether models involving those features perform well in therapeutic environments. We also excluded studies that exclusively coded patient behaviour, however many patient behaviours (e.g., change-talk in motivational interviewing; Tanana et al., 2016) are indicators of therapist fidelity. Reviews that focus on patient indicators of quality therapy may be helpful complements to our review here. We included a broad range of helping professions to try and promote knowledge crossover between related fields; however, doing so may mean approaches described here do not generalise. The models that have been used in education or medicine might not perform equally well in other settings and vice versa. Even within the field of psychotherapy, models that work well on one therapeutic intervention (e.g., motivational interviewing) may not perform well for other interventions (e.g., cognitive-behaviour therapy).

Finally, our search may have missed some grey literature or publications in other languages. While we searched our chosen databases for grey literature, we did not systematically search other websites for potential papers to include. Similarly, while we did not exclude any full-texts on the basis of language, our search terms were in English, meaning we may have missed important contributions that were indexed in other languages. The authorship team of this systematic review are fluent in the other languages (e.g., German, Mandarin) and when automated translation tools (e.g., Google Translate) did not suffice, those authors helped with full-text screening. In the cases where our authorship team was not able to read the full-text, we got help from other members of our institute who were fluent in that language. However, we used comprehensive search terms and MeSH headings, ran the search in the major databases, did forward and backward searching, and sent enquiry emails to related researchers. Still, the techniques encompassing ‘machine learning’ with researchers around the world are often shared without peer review, so it is possible we missed some papers that may have been eligible.

Conclusions

The results of this systematic review have implications for both research and practice. While more work is needed to reveal what methods work best in which circumstances, our systematic review showed that machine learning is a promising tool for assessing treatment fidelity, promoting best-practice in psychological interventions (Bellg et al., 2004). Therefore, organisations and agencies may be able to use these methods to provide prompt feedback, conduct research, and scale up training to improve therapists’ work. We have also shown that automated methods are most likely to be accurate on session level prediction with larger datasets, the fewer number of codes and conceptually concrete codes. Finally, we provided recommendations for a minimal list of considerations when developing generalisable machine learning models for treatment fidelity. In sum, machine learning shows promise as a way of decreasing barriers to assessment and feedback for treatment fidelity. Doing so can improve scientific progress by improving the consistency of interventions being studied, but also improve service delivery, ensuring clients receive effective treatments that have been validated through rigorous research.

Conflict of Interest

The authors declare no conflict of interest.

Linking chapter | From Systematic Review to Dictionary of Motivational Phrases

In Chapter 2, I synthesised the applications of automated methods used to code helping professionals' interactions via their spoken language. My systematic review showed that automated coding methods performed better than chance, and some methods showed near human-level performance. Further, it showed that automated coding methods were mainly applied in psychotherapy, with few in medical or educational settings. The methods applied in education were used to predict simple concepts, such as the number of open/closed questions or class activity type set by a teacher (e.g., group work or teacher lecturing). The results indicated that the automated models performed better when they were used to predict the codes of a well-developed behavioural coding measures. Moreover, the performance tended to be better when the methods were trained using large datasets of annotated behaviours. However, to my knowledge, few such data sets exist for teachers' motivational behaviour (e.g., over 1 million teacher interactions). Also, I could not find a detailed and comprehensive classification of teacher motivational behaviours that could be used to code teacher behaviours. So, in Chapter 3, I used an alternative automated coding method—the dictionary method—that doesn't rely on large datasets of annotated behaviours to automatically code teacher behaviours. In the next chapter, I presented the development of this dictionary and the results on how well the dictionary ratings correlate with human ratings of teacher motivational behaviours.

Chapter 3 | Dictionary of Phrases for Teachers' Motivational Behaviour: A Self-Determination Theory-Based Study

Preface

I have not submitted this study for publication, but I plan to submit soon to the Journal of Educational Psychology (IF = 6.85).

Abstract

Background

Teachers' behaviours have a substantial impact on student outcomes. Measuring teacher behaviour, using student- or teacher-report, can be clouded by biases. In contrast, observational methods can be more objective, but are expensive and time-consuming. Recently, advances in natural language processing provided new approaches to analysing behaviour. In this study, I developed a dictionary of teachers' need supportive versus need thwarting language, underpinned by self-determination theory.

Method

I followed established dictionary development strategies, using experts to refine the dictionary. I used multiple strategies to create a pool of candidate words. Experts in motivation and education research then appraised the face validity of the words. Also, to filter the dictionary empirically using the transcripts of teachers, I split the data 70:30 into 'training' and 'test' sets, where the training set was used to create a filtered dictionary, and the test set was used to compare the concurrent validity of the filtered and unfiltered dictionary. In the training set, I compared transcripts of the most need-supportive teachers against the least-supportive teachers—as judged by observers—using the weighted log odds ratios. This method identifies the words that were more likely to be used by need-supportive teachers. I filtered the dictionary using these odds ratios, keeping the need-supportive words that were more common among the most need-supportive teachers, and the need-thwarting words common among the least need-supportive teachers. To test the concurrent validity of the dictionary, I evaluated both the full, expert-derived dictionary and the filtered dictionary against observer ratings of teacher need support using the test set.

Results

The expert-derived dictionary consisted of 227 words, including 149 words for need

supportive teaching and 82 words indicating need thwarting teaching. The correlation between dictionary ratings of need support and observer ratings of need support was moderate and significant ($r_{unfiltered\ dictionary}$ = .34). This correlation was similar to the correlation between ratings by two observers of the same lesson ($r_{filtered\ dictionary}$ = .32). Filtering the dictionary using weighted log odds improved accuracy on the training set (r = .49) but reduced accuracy on the test set, not used to filter the dictionary ($r_{unfiltered\ dictionary}$ = .73; $r_{filtered\ dictionary}$ = .64). As a result, the expert derived dictionary is likely to perform better in most classrooms.

Conclusion

This expert-derived dictionary provides a cost-efficient and reliable method for analysing transcripts of teachers' motivational language. While the validity of the dictionary would be strengthened by triangulating results against other sources (e.g., student-ratings, changes in outcomes), it appears to predict observer ratings as consistently as observers. Researchers, teaching staff and policymakers may use this dictionary for analysing teachers’ motivational behaviour at large scale. It may also contribute to the real-time analysis of teachers’ motivational behaviour for feedback and faithful deployment of theory-driven interventions.

Introduction

To study the effects of teachers' behaviour on student motivation, researchers must observe what teachers do. Students can be biased reporters of teacher behaviour. For example, students give better ratings to physically attractive teachers (Riniolo et al., 2006). Teachers also provide biased reports of their own behaviour. Famously, Cross (1977) found that 94% of university professors felt they were 'above average'. Similar patterns of self-preservation bias have been found in school teachers (Kopcha & Sullivan, 2007). These student and teacher reports are still, of course, useful as they are often the only pragmatic choice. But, to robustly assess teacher behaviour in the classroom, it is more objective to observe what they *do* (Muijs, 2006). The problem is that observation is labour intensive. Paying observers to watch and rate lessons is expensive, and the cost prohibits large sample sizes (Moyers, Martin, et al., 2005). In most domains, human judges are also noisy (Kahneman et al., 2021), meaning different people make different judgements, and the same people make different judgements on different occasions. Training and calibrating observers can help (Atkins et al., 2012; Haerens et al., 2013a), but noise is generally still a problem (Kahneman et al., 2021).

One possible solution to many of these problems are automated methods of coding teacher behaviour. Automated coding methods can process large data sets at small marginal cost, and in some cases with accuracy comparable to observer judgements (Ahmadi et al., 2021). But, creating these methods is not free or necessarily easy. One form of automated coding method is machine learning, where models are built inductively from data (Ahmadi et al., 2021). Machine learning can be incredibly powerful, but training machine learning algorithms generally requires millions of coded examples of the behaviour under investigation (Brown et al., 2020; Tanana et al., 2016). This amount of data is seldom available in education, so few large machine learning models exist to assess teacher behaviour (Ahmadi et al., 2021). One alternative approach is to use 'dictionaries' that count the number of times relevant words are used. For example, Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2015) is a text analysis program with more than 80 built-in dictionaries. It scores, for example, the percentage of words in a text that relate to Sadness (e.g., 'crying', 'grief', 'sad'). These dictionaries can be developed conceptually, ideally via a panel of experts, without needing thousands of annotated examples of each behaviour. The goal of this study is to create a dictionary to automatically code motivational teacher behaviour, underpinned by self-determination theory (Ryan & Deci, 2017).

Teacher Behaviour and Self-Determination Theory. Teachers play a key role in students' motivation, engagement, and achievement. Recently, reviews on student outcomes highlighted the importance of teachers' motivational behaviour (Bureau et al., 2022; Lazowski & Hulleman, 2016) and the impact on student outcomes (Howard et al., 2021). Self-determination theory (SDT, Ryan & Deci, 2017) describes how teachers influence these outcomes. A core tenet of SDT is that students become more motivated when their teachers satisfy their basic psychological needs for autonomy, competence, and relatedness (Niemiec & Ryan, 2009; Reeve et al., 2004; Tessier et al., 2010a; Vasconcellos et al., 2020). However, studies assessing teachers' motivational behaviour mainly relied on traditional methods such as teacher- or student-report (Cheon et al., 2012; Hagger et al., 2003; Marsh et al., 2006). Those that have relied upon observer ratings have been limited in the number of teachers they could observe (Jang et al., 2010; Reeve et al., 2004; Van den Berghe et al., 2016). Still, both questionnaire and observational studies have identified a number of behaviours that appear characteristic of more motivating teachers.

Specifically, teachers are more motivating when they use behaviours that nurture psychological needs, and avoid behaviours that thwart them. They can, for example, provide rationales and choices (i.e., support autonomy; Mageau et al., 2015), acknowledging individuals progress and providing positive feedback (i.e., support competence; Sheldon & Filak, 2008), or providing interpersonal closeness and respect (i.e., support relatedness; Vansteenkiste et al., 2010). In addition to supporting these psychological needs, motivating teachers avoid behaviours that thwart those needs (Sheldon, 2011; Vansteenkiste & Ryan, 2013). They avoid controlling language or external rewards (i.e., thwarting autonomy; Bartholomew et al., 2009), avoiding demeaning their ability or emphasising their failures (i.e., thwarting competence; Sheldon & Filak, 2008), and avoid making them feel judged, isolated or neglected (i.e., thwarting relatedness; Sheldon & Filak, 2008; Vansteenkiste et al., 2010).

Objectively identifying these behaviours is important to assess the hypotheses of self-determination theory. For example, the theory proposes that teachers influence student motivation via their support for psychological needs (Reeve & Cheon, 2021; Ryan & Deci, 2017), and it is important to assess whether those relationships hold when observing teacher behaviour directly. Interventions designed to improve motivation self-determination theory also benefit from observing these behaviours to ensure those interventions are being implemented faithfully. Finally, teachers may benefit from having these behaviours observed so they can receive feedback about where they are supporting needs, and where they might be thwarting them. So, there are benefits for teachers and researchers in finding more efficient ways of objectively assessing need-supportive (and need thwarting) teaching.

Measuring Need Supportive Teaching Using a Dictionary. Need supportive teaching is likely expressed by what teachers say and how they say it (Weinstein et al., 2018). Still, it is likely that the words people use communicate a substantial component of the meaning (Pennebaker et al., 2015), and an increasingly popular way of analysing language are dictionaries (Boyd, 2017; Iliev et al., 2015). As mentioned earlier, these dictionaries work by identifying the number of words that correspond to a particular construct (as a proportion of the text). Dictionaries have been validated across a range of constructs. For example, dictionaries have been used to assess family adjustments and conflicts (Robbins et al., 2013), depression and depression-vulnerability (Rude et al., 2004), physical health (Eichstaedt et al., 2015), stereotype content (Wang et al. 2016), agency (Pietraszkiewicz et al. 2019), power (Donohue et al. 2014), and language relevant to moral foundations (Graham et al., 2009). While tone and body language influence many of these forms of communication too (Low et al., 2020), language alone conveys enough meaning for these dictionaries to measure these constructs (Pennebaker et al., 2015). The goal, with this study, is to see whether the same is true in education. Can we measure the need support of teachers—as judged by observers—using the transcripts of those teachers' lessons?

Text analysis can be conducted using many algorithms, but dictionaries have a number of advantages. They are easy to use and interpret compared with more complex analysis methods such as topic modelling (Iliev et al., 2015; Silge & Robinson, 2017). These more complex models can, for example, include more context and meaning in their judgments (e.g., they understand 'king' and 'queen' are related in a similar way to 'boy' and 'girl'). The problem with these models is that those complex relationships require many examples to learn. For example, Google's PaLM is one of the most powerful natural language models to date, but was trained on 780,000,000,000 words (Chowdhery et al., 2022). Clearly, accumulating this number of teacher sentences would be an impossible task. Dictionaries can instead be built using the wisdom of experts. For example, the founders of Moral Foundations Theory (Graham et al., 2009) used a process of 'expansion' and 'contraction' to first brainstorm words associated with each moral foundation (e.g., 'fairness/reciprocity') and then delete words that were less commonly used in the context of morality (e.g., the word 'just' is more commonly used as meaning 'only' than the more moral-laden 'fair'). This process alone was enough to discriminate between religious sermons that were more liberal than conservative (Graham et al., 2009).

In this study, I aimed to use a similar process to identify words that are indicative of more need-supportive than need-thwarting teaching. I used the existing literature to create a long list of candidate words that may indicate need supportive teaching. I then asked experts in self-determination theory to select the words on that list most likely to indicate supporting or thwarting of each psychological need (e.g., autonomy supportive, competence thwarting). I repeated this process a second time to create a comprehensive list of words, and tested the validity of the resulting dictionary against observers who rated teachers in mathematics and physical education classes. In doing so, I aimed to create a dictionary that could be used in various educational contexts to assess how need-supportive and need-thwarting teachers are, using fast, efficient, and objective methods.

Methods

Study Design

To create need-supportive and need-thwarting dictionaries, I used a methodology based on the established approach to creating the LIWC dictionaries by Pennebaker and colleagues (2015). To refine the dictionary using the available data, I used the weighted logs odds (Monroe et al., 2008; Silge, 2022). I assessed the concurrent validity between the dictionary and observer scores using Pearson's correlations.

Data Source

Text for this study was from a longitudinal study of student engagement over the first year of high school (Year 7). The study was conducted in Australian schools with below-average socio-economic backgrounds, where disengagement and drop-out are more common (Parker et al., 2018). The study focused on physical education and mathematics as subjects with important benefits for long-term participation (García-Hermoso et al., 2021), but where disengagement is common (Barkoukis et al., 2010). For the purposes of this study, these data were ideal as they allow us to assess need-supportive language in authentic classroom environments and heterogeneous contexts (i.e., both classroom-based mathematics lessons and practical physical education lessons).

As part of the study, 94 teachers consented to having a lesson filmed during the middle of the school year (70% physical education; 30% mathematics). Each of these lessons was independently rated by two trained observers on each component of need support (autonomy support, relatedness support, competence support). For this study, each lesson was transcribed using a professional transcription service. The full corpus of 94 lessons was 610,298 words and 170,208 sentences.

Expert-Generated Dictionary

I asked experts in Self-Determination Theory and education to help generate a list of words that were likely to indicate need-supportive or need thwarting teaching. This process involved five steps:

1. initial collection of words from existing sources;
2. base-rate analysis to ensure words were used by teachers in classes;
3. expert assessment of the face-validity of each word;
4. expansion of the word list using synonyms and antonyms; and finally
5. base-rate analysis (Step 2) and expert assessment (Step 3) for the expanded list (from Step 4)

Step 1. Word Collection from Existing Sources

I collected a preliminary list of potential words for each category of the dictionary. To do this, I extracted words from existing LIWC dictionaries, questionnaire items, and examples of teacher behaviour published elsewhere.

Step 1.1—Words from LIWC Built-in Dictionaries. LIWC contains built-in dictionaries to assess different linguistic, social, cognitive and psychological constructs

(Pennebaker et al., 2015). The most recent version of LIWC (Pennebaker et al., 2015) consists of almost 6,400 words. These words have been chosen based on many criteria such as construct validity, face validity, and rate-of-use in real datasets. I started these dictionaries by reviewing the built-in dictionaries assessing conceptually similar constructs. For example, the 'negative emotion' category contained some possibly 'relatedness-thwarting words', such as 'hurt' or 'nasty.' The specific dictionaries I reviewed are available in Appendix B.1. The words I extracted are in Appendix B.2.

Step 1.2—Words from Independently Developed Dictionaries. In addition to the internal LIWC dictionaries, other authors have created custom dictionaries to assess conceptually similar constructs. To find these dictionaries, I searched Google Scholar and sent enquiries to relevant communities (e.g., the SDT LISTServ). I identified one explicitly SDT-related dictionary of autonomous versus controlling self-talk (Appendix B.3, Oliver et al., 2008). I obtained this dictionary by emailing the corresponding author. I also identified a small number of conceptually related dictionaries. I used the Affiliation and Power dictionaries developed by Donohue and colleagues (2014), which assess constructs similar to need support (e.g., 'agree', 'connect', 'friend', 'kindness') or need thwarting (e.g., 'control', 'erupt', 'take'). Similarly, I identified conceptually similar words from dictionaries assessing 'agency and communion' (Pietraszkiewicz et al., 2019), 'stress' (Wang et al., 2016), and 'stereotype content' (Nicolas et al., 2019).

Step 1.3—Words from Questionnaire Items. Questionnaires assessing motivational constructs may be a rich source of words that indicate motivational constructs. I reviewed eight common questionnaires assessing support and frustration of basic psychological needs and extracted the key words from each item. Specifically, I reviewed all items of the following questionnaires: learning climate questionnaire (Williams & Deci, 1996), Interpersonal Behaviour Questionnaire (IBQ; Rocchi et al. 2017), Psychological Need

Thwarting Scale (Bartholomew et al., 2011), Basic Need Satisfaction in General (La Guardia et al., 2000), The Psychological Need Satisfaction in Exercise Scale (Wilson et al., 2006), Basic need satisfaction (Affective Feeling Scale; Reeve & Sickenius, 1994), Basic needs support (Teacher as Social Context Questionnaire; Belmont et al., 1988), and Basic Psychological Need Satisfaction and Frustration Scale (Chen et al., 2015). Then, I extracted the key words from the item that were possible signals of need support or need thwarting. For example, from the Interpersonal Behaviour Questionnaire (Rocchi et al., 2017), I reviewed the item "When I am with people who are important to me, I support their decisions" and extracted words like 'support' and 'decision'.

Step 1.4—Words from Examples of Teachers' Motivational Behaviour. In Chapter 4 of this thesis, I conducted a Delphi study and developed a classification of Teachers' Motivational Behaviour (TMB) in education. This classification encompasses the major SDT-based teacher behaviour hypothesised to change student motivation. The classification includes the behaviour and its description, function description (i.e., the way it affects basic psychological needs), as well as an example behaviour. I used the example behaviours to extract keywords that may indicate need support or need thwarting. For example, the teacher motivational behaviour 'Student input or choice' contained an example 'You can either work with a friend or do it by yourself', so I extracted 'choice', 'choose' and 'either' for the dictionary.

Step 2. Base Rate Assessment

Some words from the above sources (e.g., from questionnaire items) may seldom be used by real teachers (e.g., 'restrict'). If dictionary words are rarely used by teachers, then they will be unlikely to provide valid information about how need-supportive the teachers are. So, to create a more parsimonious dictionary and to reduce the burden on experts, I removed dictionary words that were never used in the sample of teacher transcripts.

Specifically, I removed words from the dictionary if they never appeared in any of the 94 teacher transcripts.

Step 3. Judge Rating Face Validity

After collecting words and filtering them, I asked a group of six experts to rate the face validity of the included words (Pennebaker et al., 2015). They identify whether each word corresponded to a psychological need (e.g., 'relatedness supportive'). Experts could also add new words if they felt the word fit within a category. In the previous dictionary development studies, the criteria for inclusion of words was the agreement of two-thirds of the judges (e.g., Pennebaker & King, 1999; Pennebaker et al., 2001). So, if four of the six experts put a word in the same category, I judged it as having met consensus as an example of that category. The words which did not meet consensus as corresponding to a psychological need were removed.

Judging Panel Selection. To select the experts in the field of motivation and education, I applied the selection criteria recommended for experts in a Delphi study (Baker et al., 2006). Baker and colleagues (2006) presented common characteristics of Delphi panellists and emphasised the importance of knowledge and experience. Keeney and colleagues (2017) recommend at least three years of post-qualification experience, working in the relevant area (Keeney et al., 2017). Other researchers recommended criteria such as possessing a higher degree (Keeney et al., 2001) or having a certain number of relevant publications (Duncan et al., 2004).

In our study, I invited experts from the field of motivation and education with expertise in SDT. I invited individuals if they: 1) possessed a PhD in psychology, education or educational psychology; 2) published at least three papers in peer-reviewed journals during the last 5 years related to SDT; 3) had at least 5 years of relevant experience in educational psychology, and 4) were currently employed as a researcher or academic.

Step 4. Word List Expansion

In order to expand the dictionary words, I used the words that reached consensus in Step 3 to generate additional candidate keywords. This involved a number of strategies. Where experts agreed that a word supported a need (e.g., 'group' indicated relatedness support) I constructed antonyms as candidate words for need thwarting (e.g., 'alone', 'by yourself'), and vice versa for need thwarting words.

Similarly, I generated synonyms using two methods. I consulted two established thesauri (https://www.thesaurus.com/; http://www.roget.org/) to identify synonyms of words that met consensus. I also used a machine learning model to also identify words that were conceptually related (Wordnet; Feinerer et al., 2020).

I combined these synonyms and antonyms with the new words suggested by experts in Step 4. This combined list of new words was used in Step 5.

Step 5. Assessing Base-Rates and Face Validity of New Words

To assess whether these new candidate words were used by teachers, I assessed whether those words were used by a random sample of teachers via the methods outlined in Step 2. Then, to assess whether the new words were likely to indicate need-support or need-thwarting, I asked experts to also rate these new words using the methods outlined in Step 3.

Filtering the Expert-Derived Dictionary

As outlined in Steps 1–5 above, I used existing literature and expert consensus to develop a dictionary of words the experts felt were likely to indicate need support (and need thwarting). However, experts were making their judgments devoid of context. It is possible that in a real classroom, words conceptually related to need support may be used in need thwarting ways, and vice versa. For example, 'because' was identified as a need-supportive word, likely due to its frequent use in rationales ('Algebra is useful because…'). Providing rationales is an emblematic autonomy supportive teacher behaviour (Reeve & Halusic, 2009).

However, it is also possible that ‘because’ can be used in a controlling way (‘Do it because I said so!’). This section of the method was designed to filter the dictionary so it better predicted need-supportive and need-thwarting teachers, triangulated using other data. To do this, I used the weighted log-odds method because it was more reliable and valid for real textual analysis, compared with alternatives like Cronbach’s alpha and ‘Term Frequency-Inverse Document Frequency’ (TF-IDF).

Cronbach’s alpha is commonly used in questionnaire design to assess internal consistency (Tavakol & Dennick, 2011): that is, are most of the items in this subscale assessing a similar construct? The same logic is often used for constructing a dictionary (Pennebaker et al., 2015): that is, are the words in this dictionary assessing a similar construct? For example, if a teacher is ‘competence thwarting’, then one would presume that the teacher would use many of the ‘competence thwarting’ dictionary words. This makes logical sense but textual analysis makes calculating Cronbach’s alpha problematic. For example, we can calculate Cronbach’s alpha using the number of times a teacher used a competence thwarting word (e.g., ‘wrong’; Pennebaker et al.; 2015). However, using ‘counts’ for each word means some teachers may be disadvantaged for speaking more or having longer transcripts (e.g., they may say ‘wrong’ more because they were recorded for 90 minutes instead of 60). Alternatively, we may instead control for the length of the transcript by assessing whether the word is in the teacher’s transcript at all (Pennebaker et al.; 2015). This ‘binary’ or ‘corrected method’ uses the Spearman-Brown formula, but it is also sensitive to the length of the transcript for rare words (e.g., if ‘wrong’ is rare but random then the 90-minute transcript is more likely to contain it; Pennebaker et al.; 2015). In addition, this method does not easily discriminate between more and less frequency uses of the words (e.g., ‘No! This is Wrong! Wrong! Wrong!’).

One solution to this problem is to use a measure known as Term Frequency–Inverse

Document Frequency (TF–IDF; Ramos, 2003). This method calculates how frequent a word is within a teacher's transcript ('term-frequency'), divided by how common the word is across all teacher transcripts ('inverse document frequency'). For example, a high TF-IDF for 'wrong' would mean that this teacher uses the word 'wrong' *relatively* frequently: more frequently than other teachers use the word 'wrong'. The problem with this metric is that common words (e.g., 'good') can seldom be used to discriminate between categories (e.g., need-supportive and need-thwarting language) even if the word is more common among one category (e.g. 'competence supportive' teachers). That problem occurs because the formula takes the logarithm of the percentage of teachers who use a word. If all teachers use a word (e.g., 'good') then the word's TF-IDF score will become zero (i.e., log(1) = 0). If the word 'good' was indicative of a 'competence supportive' teacher, then using TF-IDF would be ineffective for abbreviating a dictionary, because the method would remove most words commonly used by teachers.

One solution to these problems involves using the 'weighted log odds' (Monroe et al., 2008; Silge, 2022). This method starts with a prior that each word (e.g., 'good', 'wrong') are equally used by each group (e.g., need supportive and need thwarting teachers; Monroe et al., 2008; Silge, 2022). The weighted log odds method then constructs a model to assess whether usage of the word (e.g., 'good') is more/less common in this teacher, compared with the rest of the teachers (Monroe et al., 2008; Silge, 2022). By extension, the method can be used to see whether a group of transcripts (e.g., those of 'need supportive' teachers) differ from another corpus (e.g., transcripts of 'need thwarting' teachers). The method controls for the variance between transcripts by calculating a z-score of its log-odds ratio (Monroe et al., 2008; Silge, 2022). This makes the resulting metrics interpretable: positive scores indicate a stronger tendency for the transcript to include the word; negative scores indicate a lower tendency.

For these reasons, I used the weighted log odds method to filter the dictionary using observer ratings of need support. Compared with student reports, Muijs (2006) argued observer ratings are more objective, and observers can be trained in theoretical concepts such that they can better evaluate whether or not they are present. However, given the small sample of transcripts, I wanted to avoid using the same teachers for filtering the dataset and for testing concurrent validity. That is, to avoid overfitting (Cawley & Talbot, 2010; Yarkoni & Westfall, 2017), I split the teacher data 70:30 into training and testing, where 70% of the data was used for filtering the dictionary, and 30% of the data was used for comparing the filtered and unfiltered dictionaries. Splitting the data into training/testing partitions increases generalisability of results (Yarkoni & Westfall, 2017). The process increases generalisability because the testing sample attempts to replicate how well the dictionary would perform on new, unseen data (Cawley & Talbot, 2010; Yarkoni & Westfall, 2017). Without this process, it would be more likely that patterns found in this data (e.g., accuracy gains from the filtered dictionary) do not generalise to other contexts, because the same data would have been used to filter the dictionary and test the effects.

Among the 70% training sample, I used observer ratings to split the teachers three ways: most need supportive teachers (top 30% of the training dataset, 10 PE + 10 Maths), least need supportive teachers (bottom 30% of the training dataset, 10 PE + 10 Maths), and I discarded the teachers in the middle for the purposes of filtering the dictionary. I used the weighted log odds method to assess which need-supportive dictionary words were more common among the most need supportive teachers, and which need-thwarting dictionary words were more common among the *least* need supportive teachers. Words were removed from the dictionary when their weighted log odds were contrary to the expert predictions. For example, if experts agreed a word was 'competence thwarting' but was *more* common against the most need supportive teachers, it was removed from the filtered dictionary. I calculated weighted log odds using the *tidylo* package in *R* (Monroe et al., 2008).

I only assessed the weighted log odds for dictionary words used by at least 5 teachers. This criteria was to avoid weighted log odds scores being artificially inflated by the behaviour of any individual teacher (e.g., if one motivating teacher used the word 'darlings'). Prior to calculating the log odds ratios, I stemmed the dictionary words. All words—both in the dictionary and the corpora—were stemmed to increase the ability of the dictionary to capture related words. For example, the word 'preference' was stemmed to 'prefer*' to help the software to identify words like 'prefer', 'preference', 'preferably', 'preferred', and so on. This process is a similar method to the process using the LIWC (Pennebaker et al., 2001; Tausczik & Pennebaker, 2010). I used the *SnowballC* package in *R* to stem the dictionary words (Bouchet-Valat, 2014).

Concurrent Validity of Filtered and Unfiltered Dictionary

To assess the concurrent validity of the dictionary, I compared dictionary-derived ratings of teacher autonomy support against observer assessments of need support. To calculate observer ratings of need support, I calculated a random-effect for each teacher using *lme4* in *R* (Bates et al., 2015). This random effect allowed me to get the average need support rated by each of the two observers while controlling for systematic differences between observers (e.g., if one observer was more generous).

To assess concurrent validity between need support as rated by observers and the filtered/unfiltered dictionaries, I calculated Pearson's correlations. As a reference point, I calculated the correlation between the two observers of the same lesson. For primary analyses, I used composite need scores (need support minus need thwarting for autonomy, competence, and relatedness) due to sample size limitations (i.e., there were only 28 transcripts in the test set). While need support and need thwarting are different (Bartholomew et al., 2011), they are not orthogonal, meaning a composite score involving need support minus need thwarting is common practice (Chen et al., 2015). Exploratory analyses broke this measure into each component (autonomy support, autonomy thwart, etc.) to see if each part of the dictionary functions as expected (e.g., autonomy thwarting language predicting reduced autonomy support).

Results

Expert-Generated Dictionary

As described below, I generated two lists of possible dictionary words (2,526) then asked experts to rate which need each word corresponded to. When two-thirds of experts agreed on the need, I judged the word to have met consensus, as outlined in the methods section. This led to a total of 227 unique words in the expert-derived dictionary. The full process is outlined below.

Step 1. Word Collection

In the word collection step, I collected 601 words from the LIWC dictionaries, 70 words from independently developed dictionaries, 1,089 words from related questionnaire items, and 261 from indicative examples of teachers' motivational behaviours (overall 2,021 words).

Step 2. Base Rate Analyses

I checked for duplicate words collected in Step 1 and removed them. Then, I checked that words appeared among any of the 96 transcripts; 892 words appeared in the dataset.

Step 3. Expert Rating Step

Eight experts participated as the panellists and judged whether the word fit a dictionary category (e.g., 'thwarts autonomy'). Each word was rated by six experts. As described above, I retained the words judged to fit a category by four votes out of six. In the first round of surveying, 156 words met the consensus criteria. The judging panel also suggested new words for each dictionary category.

Step 4. Word List Expansion

To ensure conceptual coverage of the dictionary, I expanded the word list using methods described above. I also added the words suggested by the expert panel. These two processes together generated 505 new candidate words.

Step 5. Assessing Base-Rates and Face Validity of New Words

As described above, these 505 new candidate words were also assessed for presence in the corpus (base-rate analysis, Step 2) and expert assessment of face validity (Step 3). Seventy-one words met consensus in this step. Overall, the final dictionary included 227 unique words (149 word for need support, and 82 for need thwarting, Appendix B.4).

Abbreviating the Dictionary Using Weighted Log Odds

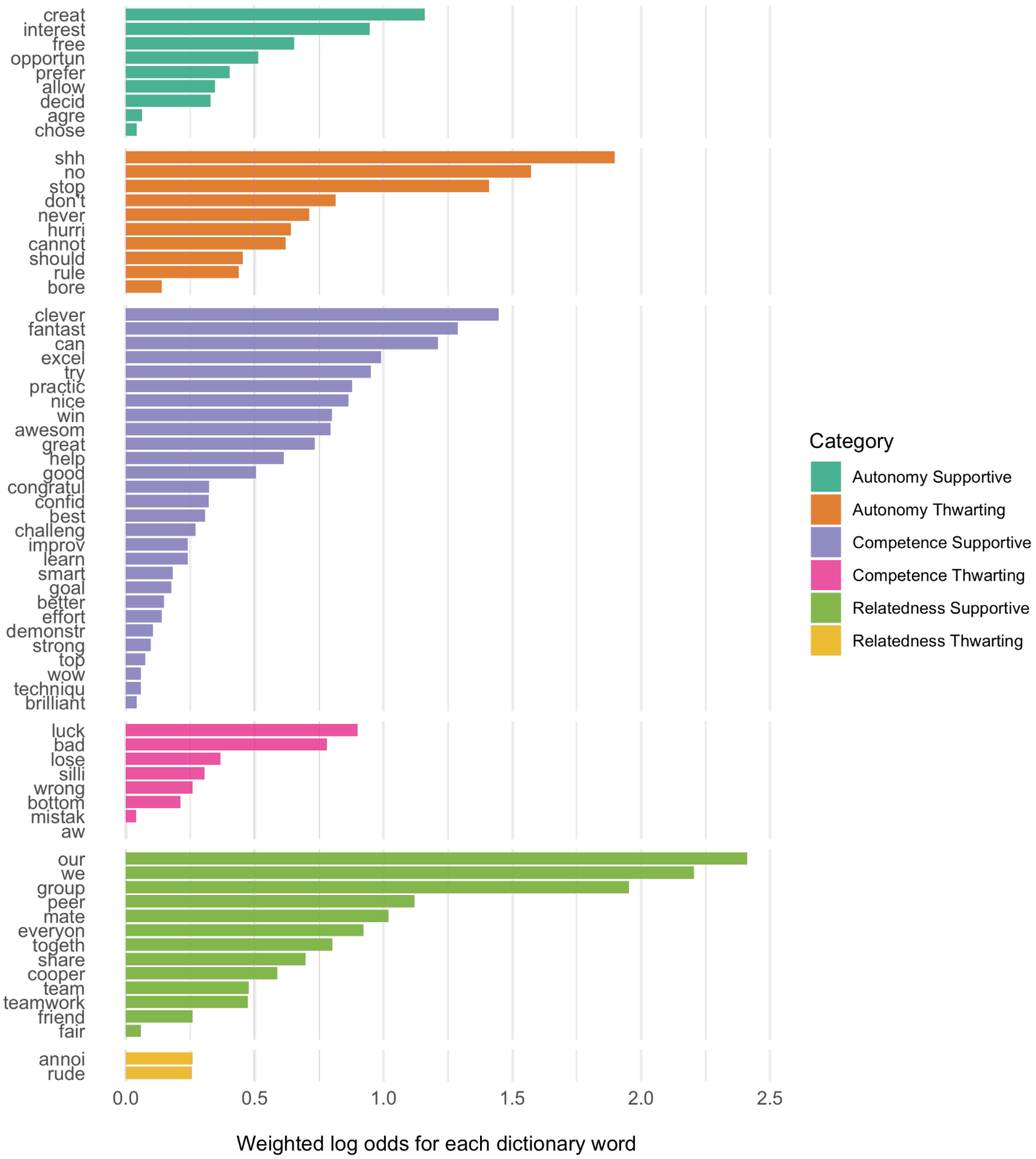
As described above, the weighted log-odds ratio represents the ability for each word to distinguish between the need-supportive and need-thwarting teachers (as judged by observers). For example, the weighted log-odds for ‘shh’ in ‘Autonomy Thwarting’ was 1.90 meaning that ‘shh’ was much more likely to appear in an ‘autonomy thwarting’ teacher’s transcript than an ‘autonomy supportive’ teacher’s transcript, as predicted by the expert panel.

Some words had negative weighted-log odds. For example, the weighted-log odds for ‘because’ in ‘Autonomy Support’ was -0.54. This means that the word ‘because’ was *more* likely to appear in an ‘autonomy thwarting’ teacher’s transcript than an ‘autonomy supportive’ teacher’s transcript. So, while the experts agreed ‘because’ was likely to indicate autonomy support (e.g., because it may indicate rationales), it may not sound ‘autonomy supportive’ to observers (e.g., “because I said so!”).

For this reason, I created a filtered dictionary that removed the words with negative weighted log odds—the words experts felt would be need supportive, but ended up more common among need thwarting teachers (and vice versa for need thwarting words). The scores for all the unfiltered dictionary words are available in Appendix B.4. I retained the words with positive log odds ratios from each category. The retained words with their weighted log odds ratio are presented in Figure 3.1.

Figure 3.1

Weighted Log Odds Ratios for Each Word in the Filtered Dictionary



The filtered dictionary consisted of 50 need supportive words (9 autonomy supportive, 28 competence supportive, 13 relatedness supportive), and 20 need thwarting words (10 autonomy thwarting, 8 competence thwarting, 2 relatedness thwarting).

Concurrent Validity of Filtered and Unfiltered Dictionary

For the concurrent validity, the main aim was to calculate the correlation between dictionary scores and observer ratings. For these analyses, I present both the concurrent validity on the 'training' data (used to filter the dictionary) and on the 'test' data (i.e., the 28 teachers not used to calculate weighted log odds scores). I used Person's correlation to calculate correspondence between the dictionary scores for a lesson and the observers' rating of that lesson.

As a reference point, I calculated a Pearson correlation to assess the inter-rater correlations between observers. These correlations were far from perfect, with moderate correlations for need support (r = .23) and small to moderate correlations when each psychological need was considered separately (see Table 3.1).

As shown in Table 3.1, the concurrent validity was comparable with this inter-rater benchmark for overall need support. Specifically, the expert-derived (unfiltered) dictionary was moderately correlated with observer-rated need support (r = .34). Observer ratings were somewhat better predicted by the need-thwarting words (r = -.37) than the need-supportive words (r = .24). More fine-grained exploration should be made with caution given there were less than 100 transcripts informing Table 3.1. Still, there was a pattern where observer ratings were uncorrelated with words hypothesised to indicate competence support (r = .16) or competence thwarting (r = -.1). On the training set, performance increased with the filtered dictionary. For example, the correlation between observer ratings and the dictionary was strong (r = .49). However, as described earlier, the training data should be interpreted with caution because the same data were used to select the words and calculate concurrent validity.

Table 3.1

Pearson correlation coefficients for the relations between dictionary and observer ratings on the training set

	Observer Ratings			
	Need Support	Autonomy Support	Competence Support	Relatedness Support
Inter-rater Correlations^	**.32***	**.23***	**.35***	**.28***
Unfiltered Dictionary				
Support – Thwarting	**.34***	.21	.37*	.34*
Need Support	**.24**	.17	.23	.23
Need Thwarting	**-.37***	-.18	-.45*	-.39*
Autonomy Support	-.07	**-.07**	-.04	-.06
Autonomy Thwarting	-.34*	**-.14**	-.42*	-.37*
Competence Support	.16	.14	**.1**	.17
Competence Thwarting	-.1	-.09	**-.08**	-.08
Relatedness Support	.25*	.14	.32*	**.22**
Relatedness Thwarting	-.26*	-.21	-.28*	**-.2**
Filtered Dictionary				
Support – Thwarting	**.49***	.39*	.46*	.43*
Need Support	**.36***	.33*	.31*	.29*
Need Thwarting	**-.43***	-.25*	-.48*	-.43*
Autonomy Support	.25*	**.31***	.14	.16
Autonomy Thwarting	-.41*	**-.23**	-.47*	-.42*
Competence Support	.2	.19	**.14**	.18
Competence Thwarting	-.12	-.1	**-.1**	-.1
Relatedness Support	.39*	.33*	.39*	**.29***
Relatedness Thwarting	-.06	-.05	-.06	**-.05**

Note: * = significant at .05 level. **Bold** involves conceptually related correlations (e.g., relatedness dictionary to relatedness observer scores). ^ Inter-observer ratings, used as a benchmark, were calculated with the full data set rather than separately for test/training.

Indeed, the data from the test set suggest that the filtering process may have ‘overfit’ to the training data. That is, it may have detected patterns in the training data that did not generalise to other teachers. Specifically, the correlations between the observer ratings and the expert-derived dictionary were higher (r = .73) than for the filtered dictionary (r = .63; see Table 3.2). Both correlations were very strong. Again, these correlations ought to be interpreted with care because there were only 28 teachers in the test-set. Still, given correlations are somewhat lower for the filtered dictionary across most constructs, results suggest the expert-derived dictionary may be more valid than the filtered dictionary in most classrooms.

Table 3.2

Pearson correlation coefficients for the relations between dictionary and observer ratings on the test set

	Observer Ratings			
	Need Support	Autonomy Support	Competence Support	Relatedness Support
Inter-rater Correlations^	**.32***	**.23***	**.35***	**.28***
Unfiltered Dictionary				
Support – Thwarting	**.73***	.51*	.74*	.73*
Need Support	**.69***	.60*	.67*	.64*
Need Thwarting	**-.37**	-.12	-.43*	-.44*
Autonomy Support	-.19	**-.26**	-.13	-.14
Autonomy Thwarting	-.39*	**-.14**	-.45*	-.45*
Competence Support	.57*	.62*	**.48***	.48*
Competence Thwarting	.06	.08	**.07**	.03
Relatedness Support	.50*	.31	.54*	**.51***
Relatedness Thwarting	.00	.09	.02	**-.09**
Filtered Dictionary				
Support – Thwarting	**.64***	.43*	.64*	.67*
Need Support	**.61***	.51*	.57*	.59*
Need Thwarting	**-.40***	-.14	-.45*	-.47*
Autonomy Support	.01	**-.1**	-.02	.13
Autonomy Thwarting	-.41*	**-.16**	-.47*	-.48*
Competence Support	.54*	.55*	**.48***	.47*
Competence Thwarting	.08	.1	**.08**	.04
Relatedness Support	.39*	.21	.41*	**.43***
Relatedness Thwarting	-.14	-.18	-.1	**-.12**

Note: * = significant at .05 level. **Bold** involves conceptually related correlations (e.g., relatedness dictionary to relatedness observer scores). ^ Inter-observer ratings, used as a benchmark, were calculated with the full data set rather than separately for test/training.

Discussion

In this study, I created an SDT-based dictionary of teachers' motivational behaviour.

This expert-derived dictionary showed moderate correlations between dictionary ratings of teacher need support and observer ratings of need support. This is promising because the relationships were as strong as the correlations *between observers themselves*. Observers are generally considered to be more objective reporters of teacher behaviour than either teachers or students, and they can be trained to attend to important, theory-driven behaviours (Muijs, 2006). But, observational assessment is expensive, and as seen in this study, noisy (Kahneman et al., 2021). That means one observer's assessment of a teacher's need support differs from another. While the expert-derived dictionary we created may not replace all the benefits of an observer, it could be a less noisy method of assessing teacher need-support at a fraction of the cost.

Another aim of this study was to see whether weighted log odds could help make the dictionary more reliable and valid, by, for example, eliminating 'need supportive' words that were used more frequently by need thwarting teachers. Toward this aim, I achieved mixed success. In the training data set, I was able to apply weighted log-odds to find an abbreviated list of words that corresponded to each need. In that training data, using only the filtered dictionary led to an increase in the correlation between dictionary and observer ratings of need-support. However, my study also applied many of the best practices in machine learning, described in my systematic review (Ahmadi et al., 2021). One of those practices include separating the data into a training set (for building a model) and a test set (for assessing how well the model does on new data; Yarkoni & Westfall, 2017). This study was a good example of why having separate training and test dataset was important. Although the abbreviated dictionary performed better on the training data, that may be because it was the same data used to create the model. On the test data, which was not used to abbreviate the dictionary, the full dictionary performed better. The filtered dictionary may have 'over-fit' to the training data, meaning it would work on the training dataset and be unlikely to generalise to unseen data (Cawley & Talbot, 2010; Yarkoni & Westfall, 2017). As a result, in most classrooms, I would recommend the full dictionary to assess need-support provided by teachers.

It is promising that a dictionary can assess teacher need support, and extends previous work showing dictionaries can assess important psychological phenomena. This study adds to the list of dictionaries for measuring psychological phenomena, including individuals' core values (Boyd et al., 2015), personality (Yarkoni, 2010), agency (Pietraszkiewicz et al. 2019), power, affiliation, and trust (Donohue et al. 2014), and moral foundations (Graham et al., 2009). These dictionaries work because a core way in which we communicate is through the words that we use. While other components of communication are important (e.g., tone; Weinstein et al., 2018), the language we use is a critical channel for communicating meaning. Devoid of any tone and context, sentences can communicate support for each psychological need: 'I care about you' (relatedness), 'You're doing well' (competence), and 'This is important for what you care about' (autonomy). As a result, it may not be surprising that experts were able to identify words that teachers might use to communicate need support.

Practical Applications

A dictionary like this may be practically useful for teachers and researchers. For teachers, a dictionary may provide feedback to teachers for how well they are using need supportive language, and how that skill improves across time. Apps already exist for providing teachers with some feedback on their teaching (e.g., talk speed, number of questions asked of students), but this dictionary may allow such apps to provide richer, theory-driven feedback about how well teachers are supporting psychological needs. Similarly, researchers have shown psychological need satisfaction is an important predictor of motivation, and so often design interventions to improve need satisfaction (Reeve & Cheon, 2021). Both observational and intervention research may be facilitated by this dictionary. Observational research may use lesson observations to more reliably assess what teachers are doing in the classroom (Muijs, 2006), and this dictionary may facilitate that research by scoring the lesson more quickly, cheaply, and reliably. Intervention research also often requires assessment of fidelity to motivational behaviour guidelines to provide stronger evidence of the causal model (Prowse et al., 2015). That is, train-the-trainer interventions are hypothesised to change teacher behaviour, and changes in that behaviour are what is hypothesised to lead to better student outcomes. Normally, assessing intervention fidelity is expensive and time consuming, and this dictionary may provide an automated method of assessing fidelity, which is much less common in education than in clinical psychology (Ahmadi et al., 2021).

Limitations and Future Directions

One obvious problem with this study was that the observer ratings appeared noisy, but those ratings were used as the reference point. It is not uncommon for people to systematically differ in their judgements (pattern noise) or for people to make different judgements on different days (occasion noise; Kahneman et al., 2021). But, with two much noise between observer judgements, those judgements become weak assessments of the underlying construct (Kahneman et al., 2021). For this study we used observers as the reference point against which the dictionary was evaluated for concurrent validity. It is possible that those judgements were insufficiently reliable and valid to be used as a ‘ground truth’ of the teachers’ need support. These observers were trained by a senior member of the research team and were not allowed to start scoring without meeting calibration criteria with that researcher. However, across domains it is not uncommon for raters to demonstrate a range of biases that are hard to eliminate, or to ‘drift’ from their calibrated point across time (Kahneman et al., 2021; Wendler et al., 2019). Future studies may want to provide more frequent opportunities for re-calibrating observers, provide observers with even more structure for maintaining consistent and valid judgements, or provide them with suggestions to avoid psychological biases.

One trade-off with these suggestions is that they require more effort for each observer, meaning the sample size may decrease for the same cost. But, sample size is already a limitation for this study. With less than 100 teachers, it was difficult to have enough power to conduct more complicated analyses without increasing the risk of multiple comparisons. These small samples may also account for the high variability in correlations (e.g., some *r*s > 0.65), especially for the test set (containing 28 teachers). With a larger sample, researchers may be able to get more stable estimates of the reliability and validity of the dictionary. Similarly, a larger sample may allow researchers to better assess each component of the dictionary (e.g., 'competence supportive' words). This may be important because, in our sample, some of the constructs (e.g., 'competence thwarting') did not appear well-measured by either the expert-derived or filtered dictionary (*r*s < 0.1). Larger sample sizes may allow for more fine-grained analysis of the dictionary performance. One method of increasing the sample size would be to use smaller 'units of analysis.' For example, if each lesson was 'sliced' into 10-minute slices and rated more frequently, then researchers could test the dictionary against each 'slice' rather than having to use each lesson as a whole. Similarly, many models in psychotherapy involve coders rating each individual sentence (Ahmadi et al., 2021; e.g., see Tanana et al., 2016). This allows those researchers to amass much larger data-sets (>1,000,000 sentences, instead of <100 transcripts), and therefore, construct more complicated machine learning models. In summary, future research may want to increase the sample size they used for training and evaluating the models they create via either recruiting more teachers or by using smaller units of analysis.

The final major limitation of this study was that the observer data was not compared against other methods of assessing need support. For example, future studies could ask students to rate the need support from their teacher after the lesson, or at the end of the 10 minute slice. Similarly, I did not assess the predictive validity of the observer/dictionary methods. According to self-determination theory, teachers who provide need support should have students who report higher satisfaction of those psychological needs, higher motivation, higher engagement and academic performance (Reeve & Cheon, 2021; Ryan & Deci, 2017). As a result, an important step in validating the dictionary is to assess whether teachers whose lessons score well end up with students who are more motivated and engaged. While beyond the scope of this study, longitudinal designs would build confidence in the assessments provided by the dictionary.

Conclusions

Dictionaries are simple and often effective methods of automatically assessing psychological variables present in text. In this study, I had experts in self-determination theory develop a dictionary of need supportive language in education. This dictionary performed well at estimating the ratings of need support provided by observers; it was as reliable as the observers were compared against each-other. I used advanced methods of attempting to filter the dictionary. My robust research design assessed whether the filtered dictionary would likely perform better or worse in authentic settings. Based on those analyses, it appears the expert-derived dictionary would be a reliable and valid assessment of teacher need support. The dictionary is an important step in automating the assessment and feedback of motivating teacher practices. Use of this dictionary in research would help efficiently measure need-supportive teaching and assess intervention fidelity. Use of the dictionary in practice would help provide teachers with theory-driven feedback for how they are motivating. This kind of rapid and trustworthy feedback is what teachers want (Link, 2022), and may help them motivate and engage the next generation of students.

Linking chapter | From a Dictionary of Motivational Phrases to a Classification of Teacher Motivational Behaviours

The dictionary showed promising results in terms of providing an overall estimate of teachers' motivational behaviours (i.e., need supportive and need thwarting) over a session. Particularly, the dictionary ratings of teachers' motivational behaviour moderately correlated with observer ratings nearly as well as the observers correlated with each other. This means that the dictionary method can be used to provide an immediate and overall rating of teachers' motivational behaviours. Thus, this method can help the research in the field by significantly reducing the time and financial resources needed for observer codings. However, the dictionary method is a simple text analysis method that relies on treating the whole transcript as an unordered 'bag of words', so it ignores context. Further, it only provides an overall impression of teachers' motivational behaviours and fails to provide detailed feedback. This method does not provide fine-grained analysis of teacher behaviours; for example, given that providing students with choices in an important autonomy supportive intervention, how often did the teacher provide the students with choice? In addition, my systematic review showed that more advanced automated coding methods perform better when they predict the codes of a reliable and consistent behavioural coding measure (e.g., the Motivational Interviewing Skills Code). However, such a coding measure for teachers' motivational behaviours has not yet been developed. In the next chapter, I developed a classification of teacher motivational behaviour that would allow the annotation of teacher motivational behaviour at a more fine grained level, while also providing a range of other benefits to the research community (e.g., better replication and translation of self-determination theory interventions). Doing so would prepare the platform for automated methods to code teacher motivational behaviours.

Chapter 4 | A Classification System for Teachers' Motivational Behaviours Recommended in Self-Determination Theory Interventions

Preface

This chapter is under second review in the Journal of Educational Psychology (IF = 6.85, SJR = 2.62). I was the first author and contributed the majority (55%) of the work (see Research Portfolio Appendix). I have retained most of the language and text as submitted. I made some minor text changes for the context of this thesis for tables, figures, and references to appendices rather than online only supplementary material.

Abstract

Teachers' behaviour is a key factor that influences students' motivation. Many theoretical models have tried to explain this influence, with one of the most thoroughly researched being self-determination theory (SDT). We used a Delphi method to create a classification of teacher behaviours consistent with SDT. This is useful because SDT-based interventions have been widely used to improve educational outcomes. However, these interventions contain many components. Reliably classifying and labelling those components is essential for implementation, reproducibility, and evidence synthesis. We used an international expert panel ($N = 34$) to develop this classification system. We started by identifying behaviours from existing literature, then refined labels, descriptions, and examples using the Delphi panel's input. Next, the panel of experts iteratively rated the relevance of each behaviour to SDT, the psychological need that each behaviour influenced, and its likely effect on motivation. To create a mutually exclusive and collectively exhaustive list of behaviours, experts nominated overlapping behaviours that were redundant, and suggested new ones missing from the classification. After three rounds, the expert panel agreed upon 57 teacher motivational behaviours that were consistent with SDT. For most behaviours (77%), experts reached consensus on both the most relevant psychological need and influence on motivation. Our classification system provides a comprehensive list of teacher motivational behaviours and consistent terminology in how those behaviours are labelled. Researchers and practitioners designing interventions could use these behaviours to design interventions, to reproduce interventions, to assess whether these behaviours moderate intervention effects, and could focus new research on areas where experts disagreed.

Keywords. Taxonomy, engagement, intervention design, behaviour change techniques, BCT

Introduction

Teachers' behaviour helps determine the quality of students' motivation and their engagement at school (Korpershoek et al., 2016; Lazowski & Hulleman, 2016; Reeve, 2009; Reeve & Cheon, 2021; Reeve & Jang, 2006; Ryan & Deci, 2017; Vasconcellos et al., 2020). When teachers foster high quality, autonomous motivation in their students, there are multiple behavioural, cognitive, and affective benefits (Bartholomew et al., 2018; Jang et al., 2010; Reeve et al., 2004; Tessier et al., 2010b). Autonomously motivated students are those who feel personal ownership and self-endorsement in their learning (Reeve & Cheon, 2021; Ryan & Deci, 2017). These students are more engaged in classroom activities and achieve better academic outcomes, compared with their less autonomously motivated peers (Froiland & Worrell, 2016; Gottfried et al., 2008; Howard et al., 2021; Reeve, 2009; Vansteenkiste et al., 2008). Unfortunately, student motivation often deteriorates over time and teacher behaviour plays a moderating role in this regard (Gillet et al., 2012; Gnambs & Hanfstingl, 2016; Lepper et al., 2005). That is, some teachers accelerate this decline whereas others can reverse the trend.

To harness the power of teachers to make a difference to student motivation, researchers have designed interventions grounded in self-determination theory (Ryan & Deci, 2020). Such interventions aim to help teachers foster students' autonomous motivation by learning to become more supportive of their psychological needs (for a review, see Reeve & Cheon, 2021). These teacher-focused interventions have been applied from early childhood to adult learning, across a range of subject domains, and in 17 different nations (Reeve & Cheon, 2021). These interventions usually comprise multiple components, such as taking students' perspectives, offering meaningful choices, and offering rationales (Cheon et al., 2012; Reeve et al., 2019). Yet, it is often difficult for readers of the subsequent publications to identify what components were used in an intervention, which component was most effective, or what each component represents in practice (Craig et al., 2008; Lazowski & Hulleman, 2016; Rosenzweig & Wigfield, 2016). This happens because intervention programs may contain different components, components may be incompletely reported, or the same components may have been labelled differently (Michie et al., 2011; Michie, Fixsen, et al., 2009). These problems present barriers to implementation, replication, and synthesis of scientific evidence. Without a good classification system of teacher motivational behaviours, it is difficult for primary research to replicate effective interventions, for secondary research to synthesise the effectiveness of such interventions (e.g., reviews and individual participant analyses; Higgins et al., 2021), and for practitioners to implement those interventions faithfully (Moreau & Gamble, 2020). As a solution to these problems, classification systems for intervention components are common practice in health and medicine where they serve to increase the quality of interventions and research (Michie et al., 2011; Teixeira et al., 2020). Yet few classifications of intervention components exist in educational psychology, potentially exacerbating failures to replicate intervention effects (Plucker & Makel, 2021). To address this gap and facilitate implementation, reproducibility, and synthesis, in this study, we created a classification system for teachers' motivational behaviour informed by SDT.

Behavioural Classification Systems Facilitate Implementation, Reproducibility, and Synthesis

In the health domain, classification systems provide a range of benefits that we aim to reproduce in educational research. Classification systems facilitate reproducibility because they provide a reliable and clear system for identifying and describing specific intervention components (Michie et al., 2011; Teixeira et al., 2020). The most useful classification systems are developed through iterative consultation with experts (e.g., Michie et al., 2013; Teixeira et al., 2020). These consultations help craft descriptions on essential components of each behaviour while trying to avoid ambiguity and confusion. It is critical to clearly understand interventions components so researchers and practitioners can reliably evaluate and implement those interventions. For example, feedback is influential in health and education (Wisniewski et al., 2019), but the kind of feedback matters. Where study authors might merely say ‘participants were given feedback on their progress’, health behaviour change taxonomies help distinguish between feedback on behaviours (e.g., step-count), feedback on outcomes (e.g., weight), biological feedback (e.g., heart rate), self-monitoring as a form of feedback (e.g., pedometers), and monitoring by others but without feedback (e.g., attendance data). Each of these types of feedback appears to have different effects for self-efficacy and behaviour, which often further varies depending on the population (e.g., Ashford et al., 2010; French et al., 2014). Classification systems help reproducibility because they allow researchers to describe interventions in a way that lets other researchers replicate the core components of the intervention (Michie et al., 2015; Michie, Fixsen, et al., 2009).

An obvious extension of this benefit is implementation. If researchers identify an SDT-based intervention that works, then practitioners working with teachers will need to know what core components were involved in that intervention. It is easier, for example, to implement an SDT intervention that specifically targets five behaviours from a clearly described list, than it is to implement a loosely defined SDT intervention without reference to specific behaviours. Classification systems can go into more detail about intervention components than is usually presented in research papers. Teixeira et al. (2020) identified detailed descriptions of SDT intervention components in health, and they explained how each intervention component supported each psychological need. If the same were available for education, it would help teachers to translate effective interventions into practice, particularly when they are less familiar with the details of the psychological theory. Although a nuanced and sophisticated understanding of the theory would be ideal, a clear and robust translation of that theory into practice could help act as a bridge between researchers and educators.

Another benefit of behavioural taxonomies is for use in evidence synthesis, like systematic reviews and meta-analyses on the effects of SDT-based interventions. Meta-analyses in education are plagued by unexplained heterogeneity (de Boer et al., 2014). Even after controlling for many features of the intervention, some interventions work better than others. The same is true in health research, where taxonomies of behavioural components have helped to disentangle some of that heterogeneity (e.g., Ashford et al., 2010; French et al., 2014; Michie et al., 2009). By being able to reliably code each intervention for the techniques that they employed, researchers can meta-analytically assess whether effective interventions are more likely to use some components, compared with the ineffective interventions (Ashford et al., 2010; French et al., 2014; Michie et al., 2009). For example, in over 100 trials to change diet and exercise, interventions that asked participants to monitor their own behaviour were more effective than those that did not, controlling for all other intervention components (Michie et al., 2009).

These kinds of conclusions are difficult to assess through individual studies because that would involve randomly assigning each possible component to see the effects on its own. Such an undertaking would be expensive and complicated. Instead, a classification of motivational behaviours would allow those involved in evidence synthesis to assess whether interventions are more effective when they employ specific intervention components. By creating a detailed classification system that experts agree upon, those doing meta-analyses are more likely to include important intervention components (e.g., to assess for the provision of choices), to code components reliably (e.g., what ‘choice’ looks like in a classroom), and to use the same vernacular across meta-analyses (e.g., such that one review looking at ‘choice’ can be compared to another).

Some taxonomies of intervention components are atheoretical (Michie et al., 2013).

These are useful for making data-driven decisions about what components work when multiple theories might explain outcomes, or when theory advancement is less focal. Other classification systems are focused on a specific theory (e.g., SDT; Teixeira et al., 2020), which has a range of advantages. Most theories hypothesise a range of behaviours that lead to improvements in motivation, and a powerful test of those theories is to see whether theory-driven interventions have hypothesised outcomes (Hagger & Weed, 2019; Lazowski & Hulleman, 2016). Researchers can become much more confident in a theory if students randomised to receive a theory-driven intervention become more motivated than those who do not, especially when effects are mediated by hypothesised mechanisms. But, to test and apply a theory via interventions, it is essential to understand how the theory links to the specific intervention components (Michie et al., 2018). Otherwise, the concordance between theory and intervention can be unclear. In health settings, 'theory-driven' interventions vary dramatically in the number of theory-adherent intervention components they use (Ntoumanis et al., 2020). Also, up to 90% of 'theory-driven' interventions do not report how each intervention component relates to the theory (Prestwich et al., 2014). We are not aware of any efforts to assess this percentage in education. This is a problem because researchers may be 'testing a theory' using an intervention that is weakly aligned to those theories. Hence, a classification system of theory-adherent motivational behaviours is essential for both intervention development and theoretical advancement in education. In this study, we focus on creating a classification of teacher behaviours based on SDT.

Self-Determination Theory

SDT is a theory of motivation that has been well-established in education (Reeve & Cheon, 2021; Ryan & Deci, 2020). It contains six 'mini-theories' that together propose a causal model for how teacher behaviour influences student outcomes (Ryan & Deci, 2017). Working backwards from those outcomes, students learn more, are more engaged, and enjoy school more when motivated by more autonomous forms of motivation (Taylor et al., 2014; Vasconcellos et al., 2020). Autonomous forms of motivation are those that are more self-directed, such as learning for the inherent joy of doing an activity ("intrinsic motivation") or as a means to personally valued goals ("identified regulation"; Ryan & Deci, 2017). In contrast, students may underperform and be less happy when motivated by controlled reasons (Taylor et al., 2014; Vasconcellos et al., 2020). These forms of motivation include feelings of obligation or contingent self-worth ("introjected regulation"), and a desire to receive rewards or avoid punishment ("external regulation"; Ryan & Deci, 2017). Autonomous motivation leads to better outcomes than controlled motivation in many domains, including education. A meta-analysis of 223,209 students found autonomously motivated students are more engaged, effortful, satisfied and happy (Howard et al., 2021). They are less absent, bored, anxious, depressed, and likely to drop out of school (Howard et al., 2021). Benefits of autonomous motivation have also been shown in meta-analyses of teacher motivation (Slemp et al., 2020), leadership (Slemp et al., 2018), and health behaviour (Ng et al., 2012; Ntoumanis et al., 2020).

The benefits of autonomous motivation are so robust because those types of motivation are driven by the satisfaction of three basic psychological needs (Bureau et al., 2022; Ryan & Deci, 2017). According to SDT, all people have a need to feel effective (the need for competence), to feel connected to those they care about (relatedness), and to feel volition in and a self-endorsement of activities they undertake (autonomy; Ryan & Deci, 2017). Consistent with SDT, the aforementioned meta-analyses all showed that autonomous forms of motivation are more likely when these basic psychological needs are satisfied (Bureau et al., 2022; Ng et al., 2012; Slemp et al., 2018; Vasconcellos et al., 2020). In education, teachers who support basic psychological needs confer a range of benefits to their students (Bureau et al., 2022; Jang et al., 2016; Reeve & Cheon, 2021; Ryan & Deci, 2020;

Taylor et al., 2014). However, thwarting basic psychological needs can contribute to a range of negative consequences, including lower self-esteem, disengagement, and poor academic performance (Bartholomew et al., 2018; Reeve & Cheon, 2021; Ryan & Deci, 2020).

Unfortunately, many teachers exhibit controlling, cold, or chaotic teaching styles (Aelterman et al., 2019; Van den Berghe et al., 2013). Controlling styles are those where teachers pressure students to follow the teacher's commands, regardless of student preferences (thwarting autonomy; Aelterman et al., 2019). Cold teachers show little personal care or concern for their students (thwarting relatedness; Van den Berghe et al., 2013). Chaotic teaching styles leave students to lean on their own, leaving them feeling overwhelmed or confused (thwarting competence; Aelterman et al., 2019). Fortunately, teachers can learn how to avoid enacting controlling instructional behaviours that thwart students' basic psychological needs and instead adopt replacement instructional behaviours that support the three psychological needs (Reeve & Cheon, 2021; Su & Reeve, 2011). They can, for example, support autonomy by providing students with choices rather than mandates, or provide rationales rather than unjustified directives (Aelterman et al., 2019; Patall et al., 2017; Reeve & Jang, 2006). They might support relatedness by acknowledging and accepting negative affect rather than punishing it, or expressing interest in students (Patall et al., 2017; Reeve & Jang, 2006). They might support competence by providing specific, informative feedback and clear goals (Aelterman et al., 2019; Patall et al., 2017; Reeve & Jang, 2006). The goal of these interventions are to simultaneously reduce the risk that teachers thwart students' psychological needs while also increasing the chance that teachers support those needs (Reeve & Cheon, 2021; Su & Reeve, 2011). In doing so, they are likely to increase student motivation, engagement, and learning (Jang et al., 2016; Reeve & Cheon, 2021; Ryan & Deci, 2020; Taylor et al., 2014).

Although student motivation is influenced by many factors, such as the values of the student (Ryan & Deci, 2017), teacher behaviours have the highest leverage for interventions because they have strong effects on students while also being malleable (Reeve & Cheon, 2021; Ryan & Deci, 2020; Su & Reeve, 2011). Learning how to support psychological needs can also confer a range of benefits to educators, who can also become more motivated by learning how to better motivate others (Ntoumanis et al., 2017). Reaching a consensus on the descriptions of these teacher behaviours is critical to improve how well we assess and implement SDT interventions. A robustly produced classification system could help us understand which teacher behaviours are most influential, and enable tests and translations of those behaviours in schools.

Robust Methods for Developing Behavioural Taxonomies

When researchers have developed behavioural taxonomies in the past, there have been two broad approaches. In the first, a relatively small group of experts—usually less than 10—write a paper where they list and describe the behaviours they think are relevant (e.g., Abraham & Michie, 2008; Michie et al., 2011). This may be similar to what educational researchers have been doing informally, listing the behaviours that the authorship team believes are consistent with that theory. Although this approach is efficient, more recent taxonomies have leveraged the Delphi method as a more formal and systematic means of gaining expert consensus (Hardcastle et al., 2017; Michie et al., 2013; Teixeira, Marques, Silva, Brunet, Duda, Haerens, La Guardia, Lindwall, Lonsdale, et al., 2020). In our study, we use this robust method to develop our classification of teacher behaviours.

The Delphi method involves asking experts to iteratively and systematically answer a number of questions, ideally until they reach consensus (Brown, 1968). Between each iteration, experts see what their peers thought, and are given an opportunity to update their beliefs on the basis of those opinions (Brown, 1968). Delphi studies aim to eliminate many of the biases that often foil group decision-making processes (Powell, 2003). For example, researchers using the method tend to assemble a large number of experts (usually > 20) to more reliably leverage the ‘wisdom of the crowd’ while aiming to maintain high standards for panel membership (Baker et al., 2006). This larger number of experts is more likely to fully cover the ‘landscape’ of perspectives on the question. Researchers using the method often de-identify the contributions of each group member so arguments are judged on their merit rather than on the personal identity of who makes the argument (Moore, 1987). They also ask for independent opinions in parallel so assessments are less likely to be clouded by the judgments of others. Applied to behavioural taxonomies, the Delphi method is likely to lead to a more reliable, clear, exhaustive, and authoritative list of behaviours than taxonomies developed by a small authorship team using ad hoc procedures (Hardcastle et al., 2017; Michie et al., 2013; Teixeira, Marques, Silva, Brunet, Duda, Haerens, La Guardia, Lindwall, Lonsdale, et al., 2020).

Aim of the Present Study

In this study, we used a Delphi method to create a classification of teacher behaviours consistent with SDT. As per previous Delphi studies that catalogue intervention components (Hardcastle et al., 2017; Michie et al., 2013; Teixeira et al., 2020), we first searched the literature to create an initial list of candidate behaviours. Next, we assembled a large group of researchers with expertise in SDT applied to educational settings. We then used the Delphi method to work with these experts to:

- clarify the descriptions of each behaviour,
- rate the relevance of each behaviour to SDT,
- align each behaviour to a basic psychological need, and
- estimate the average likely effect of those behaviours on student motivation.

The experts were also asked to identify redundant behaviours, and suggest missing ones. The ultimate goal of the process was to create a mutually exclusive and collectively exhaustive list of teacher behaviours that support or thwart psychological needs. In doing so, we aimed to create a classification system of motivational behaviours that researchers and practitioners could use to better implement, reproduce, and synthesise interventions for improving student motivation.

Method

Similar to the procedure in the previous classification systems, we applied a three-round Delphi procedure (Michie et al., 2013; Teixeira et al., 2020). For most questions, three rounds of the Delphi method are generally enough to reach an equilibrium where future rounds substantially do not change results (Delbecq et al., 1975). As described below, we assembled a panel of experts in SDT in education, generated an initial list of teacher behaviours, and used three Delphi rounds to refine that list.

Participants

To solicit diverse but authoritative perspectives on how teachers support and thwart students' basic psychological needs, we assembled a panel of international experts. In this study, we invited researchers if they:

- had a PhD in motivation, education, or applied psychology;
- published at least three articles focusing on SDT—at least one of which was an intervention—in peer-reviewed journals indexed in PubMed or Scopus in the preceding 5 years; and
- had at least 5 years of related experience in education as an academic or a researcher

These criteria are consistent with recommendations for objectively and consistently operationalising expertise (Baker et al., 2006). There are no agreed-upon standards for a minimum panel size (Jorm, 2015; Powell, 2003). As per recommendations, we used existing Delphi studies that met consensus as a guide for our sample size (Jorm, 2015). Previous studies aiming to develop a classification of behaviour change techniques recruited between

10 and 18 experts (Hardcastle et al., 2017; Michie et al., 2013; Teixeira et al., 2020). To account for the potential of attrition (Donohoe & Needham, 2009), in this study we decided on a conservative number of at least 30 experts. Expert recruitment began after the first author gained clearance from their institution's human research ethics committee.

We used recent systematic reviews to collate papers using self-determination theory interventions in educational settings (Lazowski & Hulleman, 2016; Reeve & Cheon, 2021; Ryan & Deci, 2020; Vasconcellos et al., 2020). We assessed whether the corresponding author of these papers met our criteria, and if so, we invited them to participate in our study. We also asked participants to recommend other possible experts in their networks ('snowball recruitment'). Of the 138 experts approached, 34 consented to participate (41.2% female). The participating experts were researchers with expertise in designing, conducting, and evaluating SDT-based interventions in education. There was a mix of both early-career and senior researchers (median years of research experience = 12.5; range = 5–41). The median Google Scholar *h*-index of the experts was 18.50 (range = 3–203). Most panellists also had teaching experience (median years of teaching experience = 15; range = 3–60). All 34 had experience teaching in universities (median years = 13.5, range = 1–35) and 13 had experience in schools too (of those, median years = 5; range = 1–30). The experts resided in Australia (9), USA (4), England (3), the Netherlands (3), Canada (2), China (2), Denmark (2), Estonia (2), Belgium (1), France (1), Iran (1), Norway (1), Spain (1), Switzerland (1), and Turkey (1). To assess their cultural homogeneity, we used an established measure of cultural similarity with the USA (Muthukrishna et al., 2020). By this measure, 19 panellists reported cultural identities very similar to the USA (closest 25%; e.g., Canada, Spain, Australia), 7 reported identities moderately similar to the USA (second quartile; e.g., France, Netherlands), and 7 reported identities distinct from the USA (furthest half; e.g., Iran, Philippines, Turkey, Estonia).

Developing an Initial List of Teacher Motivational Behaviours

To develop an initial list of teacher motivational behaviours, we collated behaviours from intervention descriptions, theory papers, questionnaire items, and existing taxonomies of behaviour change interventions. We screened systematic reviews for interventions and questionnaires assessing teacher behaviours (Lazowski & Hulleman, 2016; Reeve & Cheon, 2021; Rosenzweig & Wigfield, 2016; Smith et al., 2016; Su & Reeve, 2011; Vasconcellos et al., 2020). We also reviewed theory papers (e.g., Aelterman et al., 2019; Ryan & Deci, 2017) and previously-developed behaviour change taxonomies (Hardcastle et al., 2017; Michie et al., 2013; Teixeira et al., 2020). From all these sources, we collated 1,151 behaviours that could plausibly be used by teachers that might influence student motivation. We stopped when we reached saturation, that is, when all new behaviours were subsumed by behaviours already on the list.

Naturally, this process resulted in substantial redundancy, so to create a mutually exclusive and collectively exhaustive list of behaviours we used a binning and winnowing protocol (DeWalt et al., 2007; Mâsse et al., 2016). Binning involves systematically grouping things that refer to the same latent construct (DeWalt et al., 2007). Winnowing involves reducing the contents of those bins into a representative example (DeWalt et al., 2007). Binning and winnowing has been used to create a comprehensive bank of parenting practises (Mâsse et al., 2016) and patient-reported outcomes in chronic diseases (DeWalt et al., 2007). The process generally involves three steps:

1. grouping similar behaviours into bins;
2. winnowing behaviours from bins into an exemplar of that bin; and
3. refining exemplars via iterative feedback.

For Step 1, four authors created an initial list of 48 ‘bins’ for behaviours based on theory. Then, eight authors took the initial list of behaviours and placed them into those bins.

Each behaviour was classified independently and in duplicate by two of those authors. When behaviours did not fit into an existing bin, authors created a new bin, leading to an expanded list of 61 bins. For each of those bins, two authors completed Step 2—creating an exemplar of that bin. Exemplars contained:

- a meaningful name for the behaviour (e.g., “Use of pressuring language”);
- a draft description of the behaviour (e.g., “Using pressuring or controlling language when explaining tasks, providing feedback, etc.”);
- an example of the behaviour used by a teacher (e.g., "You should...", "You have-to...", "You must..."); and
- a description of the function of the behaviour in promoting or thwarting motivation (e.g., “Increases perceived external pressure to complete the task for imposed reasons.”)

This initial draft list of behaviours was then member-checked (Step 3) by the eight authors who conducted the binning, and five teachers from local secondary schools. Based on the input of these authors and teachers, two authors refined this list of behaviours before using them as the foundation of the Delphi procedure. Following this member checking, 12 motivational behaviours were added to the candidate list, meaning the Delphi procedure started with 73 possible teacher motivational behaviours.

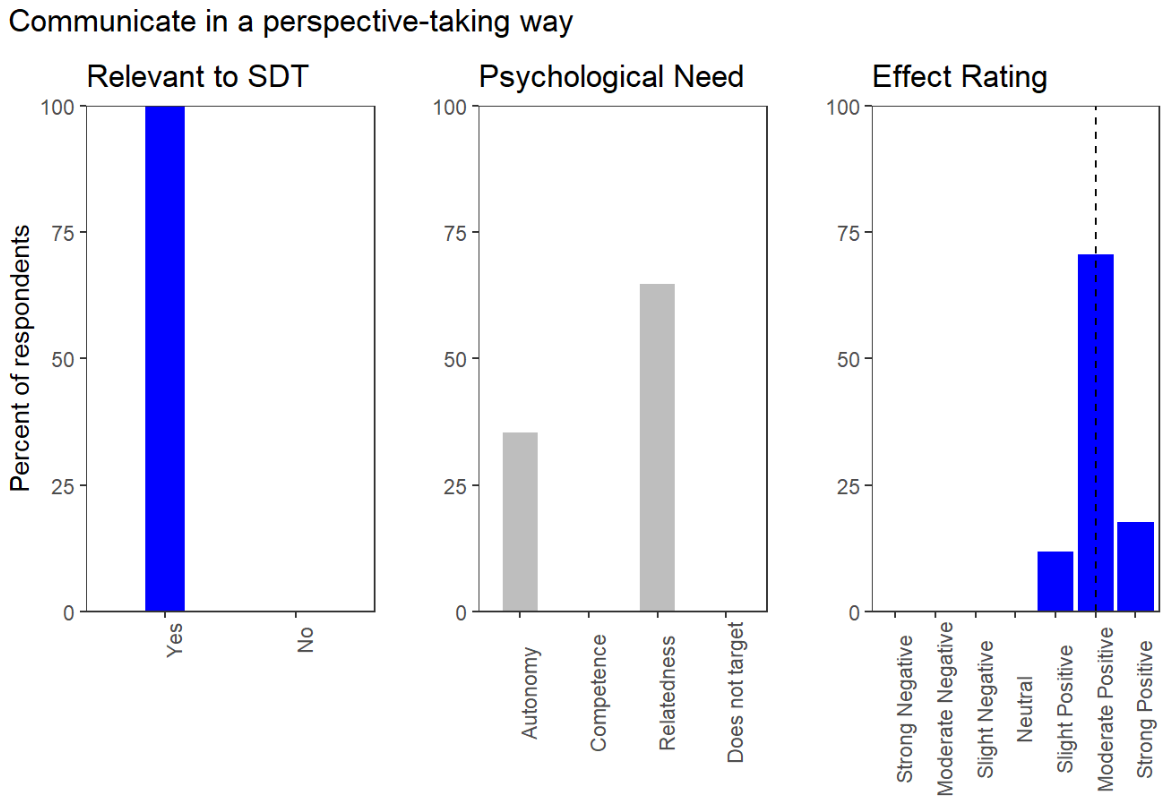
Delphi procedures

We designed and distributed the surveys online using the Research Electronic Data Capture system (REDCap; Patridge & Bardyn, 2018). In the first round, the experts provided qualitative feedback on the label name, description, example behaviour, and function description of each teacher motivational behaviour (TMB). They judged whether the behaviour was related to SDT. If their answer was yes, they identified which basic psychological need that behaviour most strongly influenced, and rated how strongly they felt the behaviour influenced motivation (7-point scale ranging from *'-3 Strong negative effect'* to *'+3 Strong positive effect'*). To help generate a mutually exclusive list of behaviours, at the end of the survey, we provided experts with a full list of TMBs and asked them to identify whether any behaviours appeared to be redundant (i.e., where two TMBs overlapped such that they described the same essential behaviour). To help generate a collectively exhaustive list, experts were also asked to nominate any other behaviours they thought were missing from the list.

After each round of the Delphi process, four authors refined the TMBs in response to the expert feedback. Where actioning recommendations involved major changes (e.g., substantially different function description), the revised TMB was considered a new behaviour, and we discarded existing ratings (e.g., of effect). In Rounds 2 and 3, we provided experts with the updated list of behaviours where ratings were available, and gave them visual feedback of the panel's responses to the previous round via bar charts (see example in Figure 4.1). Visual feedback like this helps panellists quickly see the responses of the other experts so they can assess how their beliefs compare with those of the group (Ward et al., 2014). Experts could choose to use this feedback in their updated ratings or not. Below each behaviour, we asked experts to provide qualitative feedback on the behaviour's label and description, the example, and the function description. We then also asked them to rate whether the TMB was relevant to SDT, and if so, to identify the most appropriate psychological need and the anticipated effect on motivation. We also asked them to identify missing or redundant behaviours at the end of each Delphi survey. When a TMB reached consensus on all ratings and no changes were recommended, it was added to the final list of teacher behaviours and not rated again.

Figure 4.1

Example Feedback to Delphi Panellists Provided in Round 2 and Round 3



Note. We informed panellists that the blue colouring indicated a question that met consensus, and the dashed vertical line on the 'Effect Rating' plot indicated median response.

Consensus Criteria

There are no defined standards for consensus for all questions in Delphi studies (Keeney et al., 2006; Trevelyan & Robinson, 2015). This is because it is easier for all panellists to agree on a binary choice ('yes' vs. 'no') than for all panellists to provide exactly the same score on a 7-point scale. As a result, defining consensus criteria is an inherently subjective task and should account for the nature of the question and the response scale. A systematic review of 100 Delphi studies found that the percent agreement was the most frequently applied method to achieve consensus (25 studies), although a specific agreement threshold was defined in only half of those studies (Diamond et al., 2014). Among Delphi studies, the consensus criteria varies from 51% (Loughlin & Moore, 1979) to 95% (Stewart et al., 1999).

In the current Delphi study, we used the percent agreement to analyze the "Relevance to SDT" and "Psychological Need" questions because they were nominal scales. We determined the cutoffs based on existing recommendations (Keeney et al., 2006; Trevelyan & Robinson, 2015) and previous similar Delphi studies (Hardcastle et al., 2017; Michie et al., 2013; Teixeira et al., 2020). For the binary question (i.e., "Is this behaviour relevant to SDT?"), we applied a conservative agreement level of 90% as the consensus criteria. For the other nominal question ("Which psychological need does this influence most?"), we used a slightly lower consensus criteria of 80% agreement because there were more response options, and only those who answered 'yes, this is relevant to SDT' were offered this question. This remains more stringent than the approach used in previous similar Delphi studies (e.g., 75%; Teixeira et al., 2020).

We used a different criterion for the question asking experts to rate the size of the anticipated effect for this behaviour. The panellists responded on a 7-point, ordinal scale ranging from *'-3 Strong negative effect'* to *'0 Neutral'* to *'+3 Strong positive effect'*. We judged the median to be an appropriate measure of central tendency. In line with the most conservative recommendations from a systematic review of Delphi studies (Diamond et al., 2014), we defined consensus as '90% of votes within one point of the median'. For example, if the median response was *'+1 Slight positive effect'* then we said the effect rating reached consensus if 90% of experts answered between *'0 Neutral* and *'+2 Moderate positive effect.'*

At the completion of the three rounds, we collated behaviours that were overlapping, which some experts had recommended for deletion. Rather than make a unilateral decision, we asked all experts to rate whether or not those behaviours should be deleted. We presented de-identified arguments for and against deletion, if relevant, and deleted a behaviour if more than 51% of experts agreed that the behaviour should be removed.

Results

Delphi Round 1 Results

Thirty-four experts completed the Delphi Round 1 survey. From the initial list of 73 teacher motivational behaviours, 21 reached consensus across all questions in Round 1 (relevance to SDT, targeted psychological need, and anticipated effect; see the Delphi Round 1 results and plots in Appendix C.1). We applied the experts' qualitative feedback and included the 52 TMBs that did not reach consensus in the next round to be re-rated. Also, experts suggested 9 new TMBs which we added to the next survey. Also, experts substantially modified the descriptive information for 2 behaviours that reached consensus in round 1 (Allow for student input or choice, and Provide conditional positive regard). Because the modifications were substantial, we treated the behaviours as new items and asked experts to re-rate them in Round 3.

Delphi Round 2 Results

Thirty-two experts (out of 34 participating experts) completed the Round 2 survey. Of the 61 TMBs in this round, 24 TMBs reached consensus for all questions (see the Delphi Round 2 results and plots in Appendix C.2). We applied the experts' qualitative feedback and included the TMBs that did not reach consensus in the next round survey to be re-rated. We removed four TMBs after being identified by a number of authors as obviously redundant (e.g., "Unfair use of praise" was the antithesis of "Fair use of praise"). Experts suggested one new TMB which we added to the next survey.

Delphi Round 3 Results

All 34 experts completed the Round 3 survey. Of the 36 remaining TMBs, 10 reached consensus for all three questions (see the Round 3 results and plots in Appendix C.3). Thirteen behaviours reached consensus as relevant to SDT, however, they did not reach consensus for "psychological need", "effect", or both. In this round, we also presented the

TMBs that reached consensus in rounds 1 and 2, so the experts could recommend any overlapping/redundant behaviours. Twenty-two TMBs were recommended for deletion due to overlap with other TMBs. As described earlier, we asked experts to vote on whether or not these should indeed be deleted. Thirty-one experts responded (91%). Based on those votes, 17 TMBs were removed and 5 TMBs were retained (Appendix C.4). Any other behaviours removed throughout the process are described in Appendix C.5. The final classification consisted of 57 teacher motivational behaviours (see Table 4.1).

Table 4.1

Teacher Motivational Behaviours (TMBs) Derived Through Expert Consensus, Ordered by Psychological Need and Effect on Motivation

#	Teacher Behaviour	Description	Example	Function Description	Effect on motivation: Median	Effect on motivation: Mean
Autonomy supportive						
AS1	Allow for student input or choice	Create opportunities for students to meaningfully direct the activities they do in class	"Feel free to work with a friend or do it by yourself"	Allows students to choose tasks that align with their priorities and capabilities; supports the ownership of the behaviour	+2	2.32
AS2	Teach in students' preferred ways	Use knowledge gleaned about the student values and preferences to design class activities customised to them	"I know you love comics so I based today's lesson on …"	Aligns lesson activities to students intrinsic reasons for learning rather than imposing extrinsic reasons	+2	2.09
AS3	Provide rationales	Explain the reason to perform the behaviour (e.g., why an activity is important and valuable, or how it might be personally useful)	"Doing these strength exercises makes our bones stronger, giving us a healthier body." "We're starting a module on the scientific method today because it helps us understand how the world works."	Students understand why they are doing an activity, and ideally aligns the task to a student's values	+2	2.02
AS4	Allow student own-paced progress	Allow students to work independently and to solve a problem in their	"Solve the puzzle at your own pace"	Lets students manage their own cognitive load so they do not get frustrated or overwhelmed	+2	1.91

					Effect on motivation	
#	Teacher Behaviour	Description	Example	Function Description	Median	Mean
		own pace				
AS5	Rely on invitational language	Instead of telling students what they must, have to, or should do, invite students to self-initiate into learning activities	“You may want to try this…” and “This behaviour has worked for students in the past who have had this same problem, would you like to try it?”	Reduces perceived external pressure to complete the task for imposed reasons and increases the sense of ownership of the behaviour	+2	1.83
AS6	Ask students about their experience of lessons	Ask students for feedback about how classes are going; could apply to either the content of lessons or the process/learning design	“On these sheets, please write down what you liked about today's lesson, what you didn't like, and what was most unclear. Remember it's anonymous.”	Gives students a safe opportunity to suggest constructive input and shape the way classes are run, so lessons can better cater to their needs and interest	+2	1.55
AS7	Teaching students to set intrinsic life goals for learning	Help students link learning to other intrinsic life goals, like helping others, being healthy, embracing challenges, or improving the world	“Reading helps me to gain knowledge about life” or “I want to use my reading skills to read to little kids”	Students will try to understand the lessons more, become better at doing the activities, so that students can help others someday, or discover something interesting	+1	1.5
AS8	Provide a variety of activities	Provide a variety of activities in a way that keeps things interesting	Teacher regularly changes the format of the class (debates one lesson, worksheets the next), and presents content in dynamic	Reduces boredom	+1	1.36

					Effect on motivation	
#	Teacher Behaviour	Description	Example	Function Description	Median	Mean
			ways (teaches US History using Hamilton)			
AS9	Provoke curiosity	Ask a curiosity-inducing question	"Why do we always see the same side of the moon?"	Piques student interest through facilitating their exploratory behaviour	+1	1.31
AS10	Discuss class values^	Collaboratively establish the values important to display in the class, or remind students of the collaboratively derived values	"We all thought helping each other was important, so if you see anyone struggling with the activities today, see if you stop to help them through the challenging parts"	Connects the activities that take place in class with values that the student cares about	+1	1.26
AS11	Provide extra resources for independent learning	Introduce extra resources for further learning or support outside of class time	"If you want more help, remember maths club before school tomorrow." "Here are some extra problems if you want to practise at home"	Allows for self-directed learning and progress outside of class time	+1	1.12
Autonomy Thwarting						
AT1	Use of pressuring language	Using pressuring or controlling language when explaining tasks, providing feedback, etc.	"You should …", "You have-to …", "You must …"	Increases perceived external pressure to complete the task for imposed reasons	-2	-2.24
AT2	Set up activities that exclude some students	Set up activities so there are times where some students are not doing	"If you have finished the questions, just sit quietly until everyone else is	Students do not have opportunities to engage even if they want to	-2	-1.82

					Effect on motivation	
#	Teacher Behaviour	Description	Example	Function Description	Median	Mean
		anything	finished"			
AT3	Set pressuring deadlines	Allow a capped amount of time for a task, or remind students they are running out of time	"Spend 10 minutes on this worksheet; We only have a few minutes left"	Adds pressure on students to work faster and finish tasks when the teachers says to	-2	-1.53
AT4	Use praise as a contingent reward	Praise students almost exclusively when they do what they are told	Teacher says to a student "Well done!" when they do what they were told	Increases perceived external incentives for doing an activity that is favoured by a teacher	-1	-1.34
AT5	Exhibiting solutions or answers^	Give answers to problems instead of letting students figure it out	"The answer is 42"	Stifles self-directed learning and provides external locus of causality for success (i.e., from the teacher)	-1	-1.23
Competence Supportive						
CS1	Provide optimal challenge	Offer students more challenging tasks if they find it too easy, or easier tasks if they find it too difficult	"Most of you could start on question 1. If you got 100% on the homework, you can start on question 13"	Students get the right amount of challenge for them	+2	2.28
CS2	Provide specific feedback	Provide feedback that targets a specific strategy for improvement	"If you keep your eye on your attacker then you can try for an intercept, but mostly focus on marking your girl." "You might make this argument more compelling with a quote from	Clarifies path toward goal achievement	+2	2.26

					Effect on motivation	
#	Teacher Behaviour	Description	Example	Function Description	Median	Mean
			the original source.”			
CS3	Praise improvement or effort	Provides praise that targets the improvement or effort from the student	“I see some excellent hard work here, and some improvements over last week’s work, especially in these areas ...”	Affirms students progress and improvement	+2	2.10
CS4	Provide feedback aimed at improvement or effort	Provides feedback to help a student improve or increase effort	“You have only used pythagoras theorem. If you combine these two rules, it will help get that solution”	Nurtures students’ progress by providing help that moves them forward in their learning	+2	1.95
CS5	Praise specific action	Provides praise that is specific to an action or quality of the student	“This answer was very good because it showed the working out in clear steps”	Clarifies behaviours that, if repeated, lead to goal achievement	+2	1.9
CS6	Fair use of praise	Appraises a student to help him/her improve or increase effort	Complementing all three people who completed a project in specific ways	Increases sense of efficacy	+2	1.84
CS7	Set goals based on self-referenced standards	Set up activities where each student has their own goal; ideally done subtly so no one perceives this differentiation as a form of evaluative feedback	“Try to jump further than last time.” “Take your code from last week and use one or two functions you haven’t used before to make the code shorter and easier to read.”	Promotes achievable goals by calibrating them to students skill	+2	1.81
CS8	Display hope, encouragement,	Provide positive expectations for student	“I know you can do this”	Stimulates perceived ability to meet goals	+2	1.69

					Effect on motivation	
#	Teacher Behaviour	Description	Example	Function Description	Median	Mean
	and optimism	success				
CS9	Demonstrating examples	Modelling or demonstrating examples	"When throwing, see how my other hand points at the target?" "Watch me: if you divide both sides by x, like this, we can solve for y."	Provides template for student to follow	+2	1.68
CS10	Provide feedback in private	Provide corrective feedback in private	Provide feedback 1 on 1 with the student	Mitigates risk of feedback being ego-threatening	+2	1.64
CS11	Clarify expectations	Provide clear instructions	"Start with problems 4.1 to 4.4 then check your answers with me"	Provides structure so students know exactly what to do	+2	1.61
CS12	Display explicit guidance	Provide clear guidance, clear goal, and clear action plans	"To understand how volcanoes work, we're going to make a model. First, grab a test-tube, some vinegar, and some baking soda."	Enables students to clearly understand what is expected of their behaviour	+2	1.6
CS13	Ask questions to expand understanding	Questioning to expand understanding or thinking	"What other sports do we use these skills?"; "When might we use division in our daily lives?"	Fosters a deeper understanding of how knowledge fits together	+1	1.5
CS14	Self-monitoring of progress and effort	Facilitate monitoring of progress, skill level, or performance	"How would you rate your performance in the last three weeks?"	Provides opportunities for accurate self-reflection of effort and progress, promoting independent learning	+2	1.48

#	Teacher Behaviour	Description	Example	Function Description	Effect on motivation: Median	Mean
CS15	Active learning	Set up activities where all students are engaged in a learning activity	"Complete this worksheet individually to figure out how heavy the Sydney Harbour Bridge is"; "Try to make a sentence using as few of these phonemes as possible"	Allows each student hands-on practice with an activity designed to progress development of a skill	+1^	1.42
CS16	Offering hints^	Give hints to help students along without giving them the "right answer"	"It might be easier to start with this formula"	Supports the student's own learning processes. Allows students to maintain an internal locus of causality during learning	+1	1.15
CS17	Use pupils as positive role models	Highlight some students as examples for the rest of the class to follow	"John, you commented on your code very well. Can we put it on the smartboard so your friends can see it?"	Increase self-belief through vicarious experiences of success	+1^	0.62
Competence Thwarting						
CT1	Publicly present critical feedback	Provide critical feedback in public so other students can hear	Provide critical feedback in front of the class	Increases risk of feedback being ego-threatening	-3	-2.74
CT2	Criticise a fixed quality	Provides critical feedback that targets a fixed quality	"You are not tall enough", "maths is not your strength", "you are always misbehaving, you can't control yourself"	Emphasises the importance of inherent (e.g., genetic) abilities for achieving success and insinuates that a student can not grow in their learning	-3	-2.52

#	Teacher Behaviour	Description	Example	Function Description	Effect on motivation Median	Mean
CT3	Criticise losing via peer comparison	Tell students when they are not doing as well as others	"You should learn from Paula who beat the whole class"	Emphasises peer comparison for establishing a sense of competence, meaning few students experience success by being the best	-2	-2.36
CT4	Chaotic or absent teaching	Leave students without clear instructions so the class waits or is disorganised while the teacher does something else	Teacher leaves students waiting when arranging papers at front; Teacher gives up on providing feedback so checks his/her emails in class	Students do not know what they should be doing to learn and do not get any feedback or structure about how to pursue goals	-2	-2.03
CT5	Undifferentiate d challenge	The same task is set for all students regardless of their level of ability	"Try to do a lay up by using the backboard." "Let's all play this Beethoven piece to the metronome."	Given natural variation in abilities, many students may be bored and others overwhelmed	-2	-1.84
CT6	Use vague criticism	Provides vague critical feedback with no instruction on how to improve	"Come on, James, you need to do better"	Creates ambiguity regarding strategies for students to increase competence	-2	-1.74
CT7	Praise winning via peer comparison	Congratulate winners so that everyone knows who did the best	"The highest score on the exam was John"	Emphasises peer comparison, facilitating incompetence in most students, while offering a few a sense of competence from being identified as the best	-2	-1.7
CT8	Set goals where students compete against	Set up activities where the goal is to do better than other student	"Whoever completes these problems in the fastest time wins"	Provides extrinsic reasons for working hard and few opportunities for success (i.e.,	-1	-1.47

					Effect on motivation	
#	Teacher Behaviour	Description	Example	Function Description	Median	Mean
	each-other			winning)		
CT9	Grouping students on the basis of ability	Grouping is done publicly and students are put in groups based on their ability so that there are "top" and "bottom" groups	"If you got more than 7/10, join this group working on Set A. Less than 7: in this group, doing Set B. If you did not complete the homework, you are over here working on Set C"	Increases public signalling of student competence, and means students are comparing themselves to others of similar abilities	-1^	-1.21
Relatedness Supportive						
RS1	Show unconditional positive regard	Act warmly towards students, especially ones who are challenging or who find the course challenging	The teacher is kind even to one student who did a task incorrectly and another who did not complete the task	Ensures performance mistakes or behavioural misconduct are not met with ego-threatening behaviour	+2	2.24
RS2	Ask about students progress, welfare, and/or feelings	Show interest in how students are doing, both emotionally and in their mastery of content	"How are you finding this activity, John"	Shows care and encourages students to express themselves openly, so they connect with their teacher	+2	2.07
RS3	Expressing affection	Be warm and kind to students	"It is good to see you, Theresa!"	Students feel they are cared for	+2	2.03
RS4	Promote cooperation	Set up activities that encourage students to work together on tasks	"As a group, work together to figure out this problem"	Allows joint pursuit toward a goal and potentially provides each other with feedback on progress	+2	1.89

					Effect on motivation	
#	Teacher Behaviour	Description	Example	Function Description	Median	Mean
RS5	Teacher enthusiasm^	Present content enthusiastically to make things fun and interesting	"Now I think this next part of the lesson is really interesting!"	Models the attitude and energy that the teacher would like the students to demonstrate; shows interest in the material	+2	1.84
RS6	Show understanding of the students' point of view^	Try to understand how students see things before suggesting a new way to do things	"I can understand that there are other things you'd rather do after school"	Helps the student feel listened-to and understood	+2	1.82
RS7	Group students with similar interests^	Create groups in the class where students with similar values or interests can work together on problems	When studying geography, grouping musical students to look at a country's music, the sporty students to look at the country's sports, and other students to look at the country's key historical events.	Allows students to work with people—and on tasks—that match their interests and values	+1	1.42
Relatedness Thwarting						
RT1	Ignoring students	During times where attending to students would be appropriate (e.g., emotional distress, misbehaviour, active learning) the teacher maintains distance or does not direct attention to the student	The teacher ignores an upset student	Makes students feel they are not valued or cared for and that their efforts are not noticed	-3	-2.79

#	Teacher Behaviour	Description	Example	Function Description	Effect on motivation Median	Mean
RT2	Use abusive language (content)^	Calling students by hurtful names when they misbehave	Calling a student "dummie" or "moron"	Performance mistakes and behavioural misconduct are met with competence-threatening punishment	-3	-2.76
RT3	Provide punishments unfairly	Provide punishments unfairly so students who misbehave are treated unequally	Punishing only one of two students who are speaking out of turn	Means structures are perceived as unreliable and students feel incompetent in terms of their ability to behave	-3	-2.59
RT4	Yell or use a harsh tone	Teacher yells to get control of the class	Yelling such as "HEY!"; "STOP IT!"	Creates a more emotionally unstable and unpredictable environment for students, increasing fear	-3	-2.47
RT5	Provide rewards unfairly^	Provide rewards unfairly so students who are doing equally well, get different rewards	Rewarding only one of three people who all completed a task	Students feel rewards are not predictable and teacher behaviour unjust	-2	-2.41
RT6	Be sarcastic	Use sarcastic negative phrases	"Class started 3 minutes ago. Soooo nice of you to join us" Or, "It's not like what we are learning today is important or anything"	Demonstrates contempt for students; reduces student self-esteem; diminishes the student–teacher relationship	-2	-2.16
RT7	Provide conditional positive regard^	Withdrawal warmth from a student in response to poor behaviour; provide warmth and acceptance only when teacher's	"Good job! You did it the way I asked you!"	Demonstrate that attention and warmth are contingent upon meeting the teachers' expectations	-2^	-1.85

					Effect on motivation	
#	Teacher Behaviour	Description	Example	Function Description	Median	Mean
		expectations are met				
RT8	Apply fair punishments^	Provide punishments fairly so students who misbehave are treated equally	Sending both of two students out of class when they misbehave or break a rule	Ensures misbehaviour is consistently and reliably met with external contingencies	-1^	-0.42

Note. Labels marked with ^ were placed in their modal category (e.g., autonomy support) but 'psychological need' did not meet consensus. Effects marked with ^ represent median but did not meet consensus. Effects are rated between strong negative (-3) and strong positive (+3).

Discussion

In this study, we built a system for identifying and classifying SDT-based teacher motivational behaviours that influence student psychological needs. Our Delphi panel met consensus on 57 behaviours being relevant to SDT. For most behaviours, the panel reached rigorous consensus criteria for the psychological need that each behaviour targeted, the most likely effect on motivation, or both.

With this classification tool, we aimed to help the fields of education and educational psychology to reproduce, implement, and synthesise effective motivational interventions. For example, observational or experimental research could systematically assess which specific teacher behaviours have the strongest effects on student psychological needs, motivation, and engagement. Researchers who test the effects of teacher training interventions could use this classification to describe which strategies they are using or to assess and report on the fidelity and implementation of those interventions. When practitioners and policymakers implement interventions at scale, they could then refer to the classification system as a source for detailed descriptions of which behaviours were included, and why they influence psychological needs. For pre-service and in-service teachers, the classification system may be a useful guide to what ‘need supportive’ and ‘need thwarting’ teaching looks like. And, regardless of whether researchers have already described their interventions using the classification, researchers conducting evidence synthesis could assess whether these teacher behaviours systematically explain differences in outcomes. For example, conducting a moderation analysis for interventions with and without ‘student input or choice’ (AS1) would test SDT’s hypothesis that choice is a potent strategy for improving motivation, via support for autonomy (Reeve & Cheon, 2021; Ryan & Deci, 2020).

Experts Agree on Many Influential Behaviours

We do not yet have meta-analytic assessments of the effects of each TMB, but our international panel of experts provide a number of recommendations for how to nurture student psychological needs. Most teachers would intuitively understand the destructive effects of yelling (RT4), unfair punishments (RT3), abusive language (RT2) and criticism of fixed qualities (CT2). However, experts also agreed on the benefits of many strategies that might be less common practice. For example, they agreed that moderate benefits for satisfying psychological needs could be achieved by providing students with rationales (AS3), allowing for input or choice (AS1), helping students find ways of monitoring their own progress (CS14), and by showing empathy for students' point of view (RS6). Some of these strategies are not common practice, and are amenable to change, so they would be a useful starting point for interventions (Reeve & Cheon, 2021).

Experts also agreed that a range of theoretically aligned behaviours may only have modest effects in practice. For example, experts agreed that there should be only small benefits from adding variety (AS8), offering hints instead of answers (CS16), or in grouping students with similar interests (RS7). They also agreed that there should be only slight motivational decreases for setting competitive goals (CT8) or using praise as a contingent reward (AT4). The experts' opinions may be influenced by the expectation that these behaviours may less directly target core theoretical mechanisms of SDT, or may have competing forces that attenuate their effects. For example, praise as a contingent reward may be a method of exercising teacher control, but the destructive effects of contingent rewards may be somewhat offset by the benefits of praise on competence. Stronger causal data—like meta-analyses of randomised trials—would help verify the relatively weak benefits of these discrete behaviours. Until then, people designing interventions may want to consider whether it is better to target more influential behaviours.

As would be expected, the majority of our consensus opinions align with theoretical models of SDT (e.g., Aelterman et al., 2019; Reeve & Cheon, 2021; Ryan & Deci, 2020).

This classification may help practitioners translate relatively abstract conceptual ideas, like 'autonomy supportive teaching' into a list of concrete behaviours that are observable in the classroom (Table 4.2). This list supports existing conceptualisations of need supportive teaching, such as the circumplex model by Aelterman et al. (2019). That model describes eight teaching 'styles' involving relative combinations of autonomy and structure. For example, 'attuning' and 'guiding' styles both provide a high level of need support, with 'guiding' styles offering more structure and 'attuning' styles being more student-directed. Aelterman and colleagues acknowledge that their model does not directly address relatedness, however the styles implicitly describe styles with high and low levels of relatedness. For example, the 'attuning' teaching includes "accepting students' expressions of negative affect and trying to understand how students see things" (Aelterman et al., 2019, p. 498). 'Demanding', 'domineering', and 'abandoning' styles all include behaviours that, according to our classification, would reduce relatedness. Our classification builds on these styles by providing the clear behaviours that exemplify support and thwarting for each psychological need, including relatedness. This is important because Relationships Motivation Theory is a key mini-theory of self-determination theory (Ryan & Deci, 2017) and meta-analyses show relatedness predicts student outcomes, even when controlling for autonomy and competence (Bureau et al., 2022).

Table 4.2

Need Supportive and Need Thwarting Teaching: What it is, and What it Looks Like

Psychological Need	Conceptual Definition	Emblematic Behaviours
Need supportive teachers		
Support autonomy	Create an environment where students feel volition, personal ownership and self-endorsement of their learning	• Allow for student input or choice (AS1) • Teach in students' preferred ways (AS2) • Provide rationales (AS3)
Support competence	Create an environment where students feels capable of achieving their goals	• Provide optimal challenge (CS1) • Provide specific feedback (CS2) • Praise improvement or effort (CS3)
Support relatedness	Create an environment where students feel accepted, understood, and worthy of attention.	• Show unconditional positive regard (RS1) • Ask about students progress, welfare, and/or feelings (RS2) • Expressing affection (RS3)
Controlling teachers		
Thwart autonomy	Create an environment where students feel pressured to conform to the teacher's agenda	• Use pressuring language (AT1) • Threaten punishments • Use controlling rewards
Thwart competence	Create an environment where students feel incapable of achieving their goals and unsure what is expected	• Publicly present critical feedback (CT1) • Criticise a fixed quality (CT2) • Criticise losing via peer comparison (CT3) • Chaotic or absent teaching (CT4)
Thwart relatedness	Create an environment where students feel demeaned, rejected, ignored, or judged	• Ignore students (RT1) • Use abusive language (RT2) • Provide punishments unfairly (RT3) • Yell or use a harsh tone (RT4) • Provide rewards unfairly (RT5) • Be sarcastic (RT6)

Note. Shortlist of behaviours created by selecting those with mean effect ratings greater than +2 or less than -2

The consensus opinions also aligned with meta-analyses of evidence-based interventions in education. For example, experts agreed that improvement-oriented feedback improves confidence (Wisniewski et al., 2019), that teachers' relationships with students are influential (Roorda et al., 2017), that instruction should be clear to not overwhelm students (Noetel et al., 2021), and that differentiation and scaffolding help learning (Belland et al., 2017; Smale-Jacobse et al., 2019). Although many of those meta-analyses targeted learning, our experts identified each as having positive moderate effects on motivation, too. We hope the detailed list of a substantial number of effective strategies, as identified by our expert panel, helps researchers and practitioners to develop effective interventions.

Areas of Disagreement are Ripe for Future Research

It could be most useful if future related research focused on areas where experts did not reach consensus. For example, experts did not agree on the effects of some teacher behaviours, like conditional regard (RT7), fair punishments (RT8), and grouping students on the basis of ability (CT9). These behaviours are likely controversial because the functional significance of these behaviours, or their meaning to participants, may vary depending on context. Grouping on the basis of ability may facilitate differentiation (CS1), but some children might feel the grouping publicly signals that they are in the less able group, undermining competence (Saleh et al., 2005). Behaviour management may be necessary to maintain class structure (Aelterman et al., 2019), but many behaviour management strategies include fair punishments (RT8) and selective ignoring (RT7; Simonsen et al., 2008). Targeted research on these controversial areas would help researchers ascertain when these strategies work, for whom, and why.

Similarly, experts did not agree on why, for example, empathy (RS6), teacher enthusiasm (RS5), and discussing class values (AS10) improved motivation. For ten behaviours, experts agreed that the behaviour influenced motivation, but did not reach consensus on the primary psychological need. It is likely that many teacher behaviours influence more than one psychological need, because all the three needs are interdependent and complementary of each other (Reeve & Cheon, 2021; Ryan & Deci, 2017). For example, ‘abandoning’ styles of teaching are likely to thwart both relatedness and competence; ‘domineering’ ones would thwart competence, relatedness, and autonomy (Aelterman et al., 2019). Similarly, autonomy-supportive teaching interventions usually increase satisfaction for all three needs (Cheon et al., 2012; Cheon & Reeve, 2013; Reeve & Cheon, 2021), and controlling teaching often thwarts all three needs (Reeve & Cheon, 2021; Ryan & Deci, 2017). Measures of satisfaction for autonomy, competence, and relatedness routinely intercorrelate, and factor analyses reveal that they often form a higher-order need satisfaction factor (Hagger et al., 2006). As a result, it is unsurprising that so many behaviours appear to influence multiple psychological needs. If it were more important to disentangle which behaviour targeted which need, experimental data would help confirm our panel's judgements. For example, longitudinal designs with mediation models could help determine whether each behaviour influences motivation by the hypothesised psychological need.

Strengths, Limitations, and Future Directions

Our study had 34 international experts participating from 15 countries with stringent inclusion criteria and high levels of panel retention. This is a larger panel than those used to develop previous classification systems (e.g., n = 10 in Hardcastle et al., 2017; n = 18 in Teixeira et al., 2020), which meant that we were more likely to cover the breadth of opinions and expertise in the field. Still, no such panel can survey all valuable opinions—our criteria may have excluded some experts who would have provided useful, unique contributions (e.g., teachers or principals without publications in SDT). For example, many of our experts have researched the effects of teacher motivational behaviours and student motivation across diverse samples; however, our experts were largely from Western, Educated, Industrialised, Rich and Democratic countries, as with most psychological research (Muthukrishna et al., 2020). While we had panellists from diverse backgrounds including the Philippines, Turkey, Estonia, and Iran, only 20% percent of experts were from countries that were culturally dissimilar from the USA. Fulfilment of psychological needs is important in all cultures, but how those needs are satisfied is influenced by development and culture (Deci & Ryan, 2002). This means our results (e.g., the projected effectiveness of each TMB) may not generalise well to other cultures or developing countries. Even within developed countries, students from different backgrounds (e.g., different ethnic, racial, or socio-economic backgrounds) can perceive teacher behaviours differently (e.g., see Patall et al., 2018). An important sustainable development goal is for *all* children to have access to quality education and lifelong learning opportunities (United Nations, 2015). So, future research may benefit from soliciting the perspectives of more experts from diverse populations and with different backgrounds (e.g., teachers and principals without research experience), and tailoring our findings to those populations.

In addition, in order to maintain our high levels of panel retention while maintaining the breadth of teacher motivational behaviours, we had to make responding to our survey efficient. This meant we needed to remove context and nuance from our examples. For example, we could not ask experts whether anticipated effects would be differentiated by gender, age, culture, level of ability or achievement, or level of socioeconomic advantage. As a result, future studies and interventions should be aware that these individual and contextual factors may moderate intervention effects. Although our Delphi study presents the likely effect of TMBs on average, those moderating factors are not well captured by our design. Similarly, some of our experts presented arguments that the consensus opinion may not have considered (e.g., on benefits of homogenous groups; Krijgsman et al., 2020) but these arguments may have been 'drowned out' by the sheer number of contrary opinions. Finally, evaluating the effect of any individual behaviour in isolation is difficult. The effect of one single need-specific TMB may be uncertain, whereas multiple TMBs may together yield a more gestalt 'motivating style'. The effect of these 'motivating styles' may be more obvious to students than the effects of any individual behaviour. Clearly, more reliable and valid effect estimates would come from evidence synthesis of teacher and student data, moderated by contextual factors. Future researchers could assess the concordance between the expert opinions here and efforts to collate the meta-analytic data for intervention effects (e.g., Hattie, 2008).

Many interventions and reviews focus on useful behaviours teachers could adopt, but one strength of this study was that we looked at both supportive and thwarting behaviours. Although they have opposite effects on psychological needs, thwarting and supportive behaviours are not mutually exclusive in teachers, because each exert differential effects on different outcomes (Bartholomew et al., 2009; Haerens et al., 2015; Sheldon, 2011; Vansteenkiste & Ryan, 2013), and profile studies reveal that teachers can exhibit both types of behaviours to different degrees (Haerens et al., 2018). As a result, including need thwarting behaviours may help researchers and practitioners not only identify which behaviours to promote among teachers, but also which behaviours to refrain from. Preventing need-thwarting behaviours may be as important as promoting need-supportive behaviours, given both types are important for different outcomes (Bartholomew et al., 2011). Ideally teachers can swap a need-thwarting behaviour for a supportive one (Reeve & Cheon, 2021). One limitation of our study was that we did not discriminate between 'need thwarting' and 'need indifferent' behaviours, despite recent arguments for the role of need indifferent behaviours (Bhavsar et al., 2019). Indeed, many of our 'thwarting' behaviours may be better classified as 'need indifferent': Chaotic or Absent Teaching (CT4) may not actively block students' satisfaction of needs; however, the disorganisation in the class leaves students' needs unfulfilled (Cheon et al., 2019; Huyghebaert-Zouaghi et al., 2021). Future research may benefit from separating the TMBs that actively thwart psychological needs from those that are need indifferent. Similarly, researchers have assessed new candidate psychological needs, like variety, novelty, and safety (González-Cutre et al., 2020; Sylvester et al., 2018; Vansteenkiste et al., 2020). Although most of these needs do not yet meet all the current criteria for 'basic psychological need' (Vansteenkiste et al., 2020), if the new needs are added, the classification would need to adapt, too.

To our knowledge, our classification system is the first to systematically aggregate expert opinion of influential teacher behaviours in education. By building our taxonomy on a well-established theory of motivation in education, we hope this will help researchers and practitioners test and apply that theory in schools and universities. One limitation of this approach is that our classification may neglect other intervention components that are not drawn from SDT. Intervention components from other theories (e.g., achievement goal theory; Huang, 2012) are often consistent with SDT because those interventions satisfy basic psychological needs. For example, growth mindsets purportedly improve engagement due to a more stable sense of competence (Sisk et al., 2018). However, not all educational psychology intervention components are clearly aligned to SDT. For example, idealised influence from transformational leadership theory was not included in our taxonomy. There are many other factors that influence educational engagement (e.g., e-learning, parenting) and other models of motivation (Lazowski & Hulleman, 2016). While our classification system is not comprehensive for all interventions in the field of education, it has been designed to cover applications of SDT to teacher behaviour, and we hope it sets a precedent for other efforts using different theoretical models. Other taxonomies may need to be developed for full coverage of the educational psychology literature.

Although our classification was designed to be comprehensive, 57 behaviours is a considerable list. It may be challenging for researchers or practitioners to monitor all 57 behaviours in real-world settings. The same challenge faces other fields like health, where up to 93 distinct behaviour change techniques have been identified (Michie et al., 2013). We judged that it would be better to provide the full list of behaviours that experts agreed would influence motivation. By providing the raw data for these 57 behaviours (e.g., both median and mean estimates of effect), we hope researchers and practitioners can filter the list for their own purposes (e.g., choosing only ‘strong’ effects, behaviours related to only one basic psychological need, or only those that are need thwarting). For instance, if one limits the classification system to remove those behaviours with a mean score between -2 and 2, then the classification system would include a more manageable list of 20 behaviours (see Table 4.2). Similarly, we hope and expect researchers and practitioners to use this classification as but one input in their evidence-informed decision-making (Newton et al., 2020). As Newton et al. (2020) argue, educators should account for their own expertise and knowledge of the learning context (learner age, culture, background, subject being studied, etc.). For example, a teacher with astute awareness of their context might decide that 'teaching students in preferred ways' (AS2) might involve providing fewer choices to students (AS1) who instead prefer clear instructions and expectations (CS11). Similarly, allowing students input or choice (AS1) might look different for a Year 1 class (e.g., 'draw your favourite animal') compared with a university cohort (e.g., 'choose the case study that's closest to your professional goals').A thumbs up from a teacher might be 'praise' in some cultures (e.g., United States) and abusive language (RT2) in others (e.g., Bangladesh). We agree that researchers and practitioners will need to adapt the behaviours and recommendations here to the age, skill, background, culture and context of the learners they are teaching.

Conclusion

In this study we developed a classification system of teacher motivational behaviours, based on SDT. We used a best-practice three-round Delphi procedure to reach consensus from an international panel of 34 experts. The resulting classification of 57 behaviours can be used to facilitate reproducibility as it clearly describes a range of teacher behaviours commonly applied in research. The classification system facilitates application and translation by giving practitioners clear definitions of each intervention component, and estimates of how effective each component is for promoting motivation. By facilitating synthesis, reproducibility and implementation of educational psychology research, we hope this classification makes it easier for researchers to find better ways of improving student motivation, and helps practitioners apply those methods to improve student outcomes.

Chapter 5: Discussion

Review of Thesis Objectives

In Chapter 1, I outlined the theoretical background for this thesis and introduced some automated coding methods with potential for coding teacher behaviours. In Chapters 2-4, I presented three studies designed to improve researchers' and practitioners' ability to efficiently and reliably code teachers' motivational behaviour. In Chapter 2, I synthesised the current literature on the applications of automated coding methods by conducting a systematic review of automated coding methods used to analyse helping professionals' behaviour. In Chapter 3, I developed and tested an expert-derived dictionary to automatically code teachers' motivational behaviour from lesson transcripts. And in Chapter 4, I developed a classification of teachers' motivational behaviour that would facilitate fine-grained automated coding of teachers' motivational behaviour, among other benefits. Finally, in this chapter, I provide an overall discussion of the thesis with potential implications, strengths, limitations, and suggestions for future research.

In Chapter 2, I presented a systematic review of automated coding methods. In the systematic review, I aimed to identify, synthesise, and critique all the implementations of automated coding methods used to analyse helping professionals' verbal behaviour in interpersonal interactions. My systematic review found that several automated coding methods were used in psychotherapy, medical care and education. Results of 52 included papers showed that the majority of the studies used models to predict codes from behavioural coding measures, indicated in helping professionals' language. Most of the studies applied more than one automated method. The most frequently applied models were Support Vector Machine ($k = 8$), Random Forests ($k = 7$), Logistic Regression ($k = 7$), J48 classifiers (a type of decision tree, $k = 6$), followed by Maximum Entropy Markov models ($k = 5$), and Naive Bayes ($k = 5$).

Studies in the psychotherapy context aimed to predict the fidelity to a prescribed therapeutic process (k = 28, 82.3% of psychotherapeutic studies). In medical care settings, the aim was to identify clients' symptoms (k = 1), topics discussed in conversations (k = 5), or conversational patterns (k = 5). In educational contexts, studies aimed to predict the number of teacher questions (k = 5) and the type of classroom activities (e.g., discussion, lecture, or group work, k = 5). However, none of the automated methods in education assessed teachers' motivational behaviours. Most of the studies applying the machine learning models did not adhere to the best practice guidelines. All the papers reported the clinical setting, dataset details, and observational units. However, a considerable proportion of them did not report data pre-processing (46.2% of studies), hold-out 'train and test' validation method (78.9% of the studies), success criteria (e.g., mean-squared error; 65.4% of the studies), or relative importance of predictor variables (e.g., which feature is most important in predicting the outcome variable; 63.5% of the studies). For example, guidelines recommend using a subset of a dataset (e.g., 70% of data) to build the method and unseen data (e.g., 30% of data) to test the final method (Luo et al., 2016; Yarkoni & Westfall, 2017). This is concerning as without transparently reporting these processes, the machine learning models are not reproducible on future data.

Nevertheless, automated coding methods generally demonstrated promising results regarding their agreement with human coders. In some instances, they achieved a kappa between .38 and .66 (fair to excellent levels of agreement with human coders). The accuracy of models (i.e., the ratio of correctly predicted codes to the total number of predictions) was greater than 50% in all studies, and sometimes higher than 80% (e.g., Chakravarthula et al., 2015; Wang et al., 2014; Xiao et al., 2016). The findings showed that, among the models, the support vector machines performed better than the other models used in different studies, and also performed well when directly compared to other models (Carcone et al., 2019).

This review showed that the automated coding methods showed near-human level performance under some circumstances. For example, models demonstrated superior performance when studies used large datasets of annotated interactions (e.g., 1,235 therapy sessions, Goldberg et al., 2020; above 9 million words, Imel et al., 2015). Also, coding frameworks with fewer behaviours and codes representing concrete (rather than abstract) concepts led to a more precise performance. Further, studies predicting the codes of a pre-defined reliable coding measure (e.g., MISC or MITI) showed more accurate performance.

My systematic review showed that the machine learning methods performed better when studies used a behavioural coding measure to annotate helping professionals' behaviour. Most existing behavioural coding measures have been designed to measure adherence to psychotherapy manuals. The most frequently applied coding measure was the Motivational Interviewing Skills Code (k = 14; Miller et al., 2003), followed by the Motivational Interviewing Treatment Integrity measure (k = 7; Moyers, Martin, et al., 2005). Seven studies used a simple, custom coding system to code, for example, whether teachers asked questions, provided instructions, or facilitated small-group activities (Nystrand et al., 2003).

In chapter two, the systematic review aimed to review the applications of automated methods used to analyse interpersonal interactions in helping professionals. This chapter revealed that although few studies used automated coding methods in education, they were used to predict more straightforward concepts and structures (e.g., number of questions or class activity type). Automated coding methods have seldom been applied to assess teachers' motivational behaviour. My systematic review showed that currently, there are two barriers to using an automated coding method for coding teachers' motivational behaviour. First, we do not have a large dataset of annotated teachers' motivational behaviours at a fine-grained or sentence level. This is important because machine learning requires many training examples for text models, and without sentence-level annotation, researchers would need to code millions of full lessons on the same scale. Sentence level coding allows for more examples of similar constructs. Because of this barrier, in Chapter 3, I developed an automated coding method using a dictionary. This method does not need a large annotated dataset or a coding framework. My systematic review also showed that automated coding methods work best when they are applied to predict the codes of a behavioural coding measure that outlines and describes a particular behavioural. However, a coding measure of teachers' motivational behaviour has not been developed yet. So, in Chapter 4, I developed a behaviour coding measure of teacher motivational behaviour that would lay the platform for more advanced automated coding.

The results of the first study showed that automated coding methods are efficient methods in coding interpersonal interaction and are capable of replicating manual coding methods. This means that the automated methods could analyse interpersonal interactions using verbal behaviour. Also, these methods can be used as an efficient alternative method for the traditional coding methods of observational or self-report. For example, self-determination theory is a useful model of motivation in education because the theory prescribes a number of concrete behaviours teachers could implement (Reeve et al., 2021). However, identifying the specific behaviours that most influence motivation is challenging. For example, to see if teachers are faithfully implementing those strategies, researchers could ask students, but those judgements may be unreliable. Teacher behaviours, like empathy, may be subtle to the point that students are unreliable reporters. On the other hand, researchers could have human coders observe lessons and code for the specific behaviours, but that would be time and resource intensive. The methods from this thesis provide opportunities for assessing the hypotheses of self-determination theory using automated coding methods. The dictionary provided proof of concept that teacher behaviour could be automatically assessed for need support and need thwarting. The systematic review pointed to some more advanced methods of automated coding that worked in other domains that could be applied to assessing self-determination theory in education (e.g., using support vector machines). Similarly, it provided some principles to apply when designing methods of automatic coding (e.g., clear, transparent criterion). Some parts of self-determination theory meet that criterion (e.g., need supportive behaviours). For those parts, my classification would help advance the theory by letting researchers identify which specific behaviours best predict student outcomes, and with larger data-sets in the future, automatically code teacher behaviour on those behaviours. This could, for example, be used to observationally assess teacher performance, to provide feedback for teacher development, or to test interventions for fidelity to the model.

In my systematic review I found few studies that had explicitly employed dictionary methods, however my study 3 found that these methods performed as well as observers in educational settings. Future studies could explore using different dictionaries in educational settings to see if they perform as well for other constructs beyond autonomy support (e.g., transformational vs. transactional leadership, mastery vs. performance motivational climates).

In a similar vein, the systematic review found that previous literature had only coded surface level annotations in educational settings (e.g., number of questions) compared with therapy settings (e.g., classification on one of 21 codes from the motivational interviewing skills code). This appeared to be due to the few, well-established classification systems available for coding teacher motivational behaviour. In my Delphi study, I created an expert-derived classification system of teacher motivational behaviour. This classification met most of the recommendations from my systematic review, in that it operationalised clear, visible behaviours with examples. The systematic review found that automated coding was more reliable for these transparent behaviours than for behaviours requiring more subjective judgement. So, my classification system means that future researchers may be able to better annotate teacher transcripts for specific behaviours, rather than merely gestalts. Being expert derived and consensus driven, it may also facilitate pooling of datasets across studies; doing so in psychotherapy was essential to train state-of-the-art machine learning models in that domain (e.g., Tanana et al., 2016). My systematic review showed that more flexible machine learning models—like support vector machines—tend to perform better than simpler models. These more flexible models are only possible to be trained with large datasets (>100,000 examples), meaning behaviours usually must be coded at the sentence (rather than lesson) level. Most existing systems of rating teacher behaviour focus on 'lesson level' ratings. For example, the dictionary I created rates the teacher's transcript across the lesson, compared against lesson-level ratings of teachers' need support. In contrast, my classification allows for more fine-grained annotations of teacher behaviour (e.g., moments where teachers provide specific feedback) and therefore allow for training more sophisticated machine learning models.

In chapter 2, I conducted a systematic review of automated methods used to analyse interpersonal interactions. The review showed that the automated methods are capable of analysing interpersonal behaviour using verbal language and under the circumstances, they indicated a near human-level performance. This study showed that interpersonal interactions could be automatically assessed in similar contexts to education (such as motivational interviewing). Also, it showed that interpersonal interactions could be reliably replicated using automated coding methods. Thus, these findings provided the base knowledge and possibility of using automated methods to automatically assess motivational behaviour in education. So, in chapter 3, I used an automated method to analyse teacher motivational behaviour. Using the dictionary method, I was able to replicate findings from the systematic review, obtaining similar levels of reliability for the dictionary as for human coders. Compared to the reliability coefficients identified in the systematic review, the dictionary I tested was strong/moderate/weak at predicting the construct of interest. Given the patterns in the systematic review, the strength of this association is likely due to the relatively vague nature of the construct. That is: it requires more interpretation and judgement to identify if a teacher is 'autonomy supportive' than to identify if a teacher asks open-ended questions. For these reasons, I created the classification system in Study 3, to operationalise 'need support' and 'need thwarting' via specific, observable behaviours. The expert-derived classification provides much clearer descriptions of each construct (e.g., 11 behaviours that are 'autonomy supportive'), which will allow future automated coding methods to achieve even higher levels of reliability, closer to those at the top end from my systematic review.

In Chapter 3, I developed an automated coding method to analyse teachers' motivational behaviours using their spoken language. To do this, I had experts develop a set of discrete words that could indicate need supportive and need thwarting behaviours expressed in teachers' language. I used these words to develop an SDT-based dictionary of teachers' need supportive and thwarting language, including for each basic psychological need. The results showed that dictionary-based ratings correlated moderately well with observer ratings—about as strongly as observers correlated with each-other. Further, I filtered the dictionary using the state-of-the-art statistical method (weighted log odds ratio), controlling for the use of words in teachers' language expressed in real-setting classes (Monroe et al., 2008; Silge, 2022). That means I filtered the dictionary using a data-driven approach that kept the most indicative need supportive and thwarting words present in real-setting teacher language. The results showed that the filtered dictionary demonstrated superior performance on the training dataset, but the unfiltered dictionary performed better in the test dataset. This finding indicated that while the dictionary method can assess teachers' motivational behaviour generally, adhering to the best practice guidelines is critical when using any automated coding methods, as identified in Chapter 2 (Ahmadi et al., 2021).

Chapter 2 also showed that automated models performed better when using a behavioural coding measure. For example, models that predicted the MISC or MITI codes appeared to achieve higher accuracy. Such behavioural coding measures define a set of explicit behaviours present in verbal interactions. For example, the MISC recommends some conversational devices (e.g., reflections, affirmations, open questions) for practitioners to follow (Miller et al., 2003). Such clearly defined codes improved the performance of automated coding methods (Ahmadi et al., 2021). In some contexts, including the health domain, Delphi studies have been conducted that allowed experts to reach consensus on both a-theoretical and theory-driven strategies (Michie et al., 2013; Teixeira et al., 2020). However, a behavioural coding measure of teachers' motivational behaviours has not been developed yet. Without such a classification, the automated coding methods might not perform their best when analysing teachers' motivational behaviours.

In Chapter 4, I created a behavioural coding measure to improve the fidelity and transparency of interventions, and to also enable reliable coding of teachers' motivational behaviours. Using a three-round Delphi procedure, I recruited 34 SDT-based experts from 15 countries to reach consensus on teachers' motivating behaviours. This classification contained definitions of each behaviour, function description, and some indicative examples for each teacher motivational behaviour. These details are designed to make it easier for people to more reliably identify teacher behaviours consistent with SDT.

The findings of the dictionary study showed that the dictionary analysis of teachers' motivational language is as reliable as the inter-observer reliability. So, as the dictionary method can assess some psychological constructs (e.g., power), the results showed that the dictionary method could automatically assess teachers' motivational behaviour as well. This is an important result because it indicated that teachers' motivational behaviour could be reliably assessed only using teachers' language. While other factors likely matter too (e.g., tone of voice; Weinstein et al., 2018), language appears to communicate a substantial part of teacher motivational valence. As mentioned earlier, this means dictionaries may be useful for robustly and efficiently testing the tenets of self-determination theory (i.e., that training teachers in need support leads them to become more supportive, and being more supportive leads to more satisfied, motivated, and engaged students). This new method of assessing teacher motivational behaviour may also be applied in SDT-based interventions to provide fast, accurate and individualistic feedback, leading to better behaviour change.

In chapter 4, I created a classification of teachers' motivational behaviour based on SDT. While the theoretical frameworks outline the motivating and demotivating constructs (e.g., need supportive and need thwarting styles), a comprehensive list of distinct and specific behaviours has not been developed for each construct. Therefore, we identified, described and provided an example for each teacher motivational behaviour using a common language. Further, we provided functional descriptions based on SDT propositions (i.e., how a behaviour improves student motivation based on SDT tenets). This is important as this system provides an instructional guideline, so researchers can use it to design effective interventions based on SDT. Moreover, this would help teachers easily understand SDT-based motivational behaviours and apply them in an educational setting. It may be helpful for testing the tenets of self-determination theory both observationally and experimentally. Observationally researchers may see whether teachers who display the behaviours in the classification do indeed have more motivated and engaged students. Experimentally, researchers may see whether training teachers in the behaviours on the classification increase student engagement and motivation.

Strengths and Implications

Overall, this thesis aimed to test the applicability of automatically analysing teachers' motivational behaviours. The dictionary I developed in Chapter 3 showed that automated coding methods could reliably assess teacher motivational behaviours in terms of their need supporting or thwarting behaviours. In assessing teacher motivational behaviours, observer ratings have been generally considered an objective, reliable method. My thesis showed that correlations between the dictionary and observer ratings were similar to those between two observers. While the expert-derived dictionary may not replace all the benefits of observational methods, it could reliably assess big datasets of teachers' motivational behaviours, in a shorter time and with significantly less cost. The developed dictionary provides some implications for practitioners, researchers and policy makers that I outlined in the following paragraphs. My systematic review showed that the machine learning models performed better when they predicted the codes of a behavioural coding measure. Part of the reason might be that such coding measures allow coders to rate behaviours at the sentence level, making the coding more accurate and reliable. The expert-derived classification I made in Chapter 4 aimed to facilitate this process by building expert consensus around a set of important behaviours. The benefits of this classification go beyond data coding for machine learning. The classifications aims to facilitate meta-analyses to identify which behaviours matter most in interventions, enable the replication of intervention research, and promote faithful translation of interventions into practice. It will make machine learning more feasible, and in the meantime, it provides a similar set of benefits to teachers and researchers.

Providing Teacher Feedback

Providing constructive feedback for teachers is essential to improve their performance at school. Both the dictionary and the classification could be helpful tools for teachers and principals aiming to enhance the quality of teachers' motivational behaviours. The dictionary-based feedback may help teachers identify the parts they need to improve. For example, by analysing the transcriptions of one teaching session, a teacher would know to what level they are using motivating (i.e., need supportive) or demotivating (i.e., need thwarting) language. And if they use the dictionary analysis again in the future, they will notice what areas they have improved and what areas need further improvement. Similarly, the teacher motivational behaviour classification could be used for observational assessments of teachers, where peers or supervisors could identify which behaviours they use well and which they could do differently.

Providing this kind of specific and individualised feedback at a large scale could be expensive and even impossible if researchers and practitioners relied on observational methods. Our dictionary could address these obstacles by providing rapid, accurate and individualistic feedback for a large number of teachers in a fraction of the time. Further, some mobile applications already exist that provide feedback on teachers' practice (e.g., visible classroom). However, such applications only provide a certain type of feedback on their verbal language, such as their talk speed or language semantics (e.g., the difficulty level of vocabulary). If such applications encompass our dictionary analysis, they could provide richer and theory-driven feedback on how well teachers are supporting their students' psychological needs.

Faster and More Reliable Intervention Fidelity Assessments

The classification system (Chapter 4) for teacher motivational behaviours could help with intervention fidelity by both helping researchers to use a common language when describing their interventions, and by allowing observers to see whether intervention teachers are indeed using the intervention components. By having this system reach consensus from a range of international experts, it makes it more likely that different researchers could use the same classification system. This means that researchers and policymakers could more easily compare and contrast the interventions being researched, making it easier to identify the teacher behaviours that matter. The details provided in the classification may increase the reliability of coding given each behaviour has such a detailed description of what it looks like.

Similarly, the dictionary could make it easier for researchers to research teacher behaviour through observational or interventional studies. Observational studies could benefit from using the dictionary to score teachers' motivational behaviours more quickly and reliably with significantly lower costs. Also, intervention studies might need to assess teachers' behaviours before the intervention, and once or multiple times after the intervention. Our dictionary can help studies overcome the barriers to a rapid analysis by significantly reducing the time and financial resources needed for coding teacher behaviours.

Better Knowledge Synthesis and Translation

The classification developed in Chapter 4 is expected to improve teaching quality, fidelity to evidence-based interventions, and the synthesis of literature on effective motivational strategies.

The classification offers a range of motivating and demotivating (i.e., need supportive and need thwarting) behaviours that teachers can utilise in their practice. That is, to support student psychological needs, teachers could use the need supportive behaviours and decrease their use of need thwarting behaviours. The behaviours were collected and introduced through a robust methodology, with the consensus opinion of 34 SDT experts. So, the classification could be a reliable source for best practice guidelines of motivational behaviours in education. Pre-service and in-service teachers could use the classification to become familiar with the explicit behaviours that would support or thwart student psychological needs. Furthermore, it has been shown that feedback and supervision can mitigate drifts in performance over time (Barwick et al., 2012; Ivers et al., 2012; Madson et al., 2009). The classification could be a useful tool to provide feedback for teachers on the quality of their motivational behaviour. Principals can use this classification to provide feedback for teachers on how frequently and to what extent they are using the teacher motivational behaviours in daily teaching practice.

Similarly, researchers and policymakers can use our classification as a reliable source to select and combine strategies that would target students' particular basic psychological needs. The classification identifies the behaviours aligned with each psychological need and expert opinion on which ones seem to be more effective. Researchers and practitioners may want to ensure those behaviours are included in their intervention programs. Further, the classification provides the main components of an intervention such as description, function description and example behaviour for each behaviour. Once studies start using our classification, and identify the most effective strategies, it would be straightforward for future research to replicate the effective interventions (Michie et al., 2011; Teixeira et al., 2020).

Finally, systematic reviews may want to aggregate the literature on the effectiveness of each teacher motivational behaviour using the classification. Meta-analyses in education are plagued by unexplained heterogeneity (de Boer et al., 2014). Regardless of whether the studies described their interventions in line with the behaviours in the classification, the classification contains enough detail to allow researchers to identify which behaviours may have been used within interventions. This would enable meta-analytic synthesis to assess which behaviours best explain differences in student outcomes.

Limitations and Future Directions

In Chapter 3, I showed that teachers' motivational behaviour could be assessed using the developed dictionary. Similar to other contexts such as psychotherapy, we could use more advanced machine learning models to analyze teachers' motivational behaviours. However, limitations such as the lack of a behavioural coding measure and insufficient annotated data prevented me from applying such a model. Particularly, my systematic review showed that machine learning models perform better when they are predicting the codes of a well-developed behavioural coding measure. However, there was not such a classification developed in education. Instead, for the dictionary study we were required to use gestalt ratings of autonomy support from trained observers. These mean the findings of the dictionary, and the capacity of the thesis, was limited by the observer data I had available.

To address this limitation in the future, I developed this classification to lay the foundation for fine-grained coding of teacher motivational behaviour, facilitating more sophisticated machine learning models. My systematic review (Chapter 2) showed that the machine learning models performed better when they were trained with a large dataset of annotated behaviours (e.g., 1,000,000 annotated psychotherapy interactions). Future research could annotate teacher transcripts using the classification (Chapter 4) to see if machine learning models can replicate those annotations and if those annotations predict student outcomes. To expedite this process, the model could be trained 'adaptively' using machine learning (e.g., via Explosion, 2022) where observers annotate examples until the machine learning model is confident in its classification, then can focus on examples where the model is uncertain.

In Chapter 3, the dictionary ratings were consistent with the observer ratings of the same session. However, I did not have enough data to investigate if the dictionary ratings of teacher behaviours would predict the hypothesised change in student motivation and engagement. Having more data on teacher behaviours (e.g., recordings of many class sessions over an educational semester/year) would allow investigating if teacher behaviours explain student motivation and other consequent outcomes.

Finally, I acknowledge the limitations of the dictionary method in assessing teachers' motivational behaviour. The dictionary results added to the literature that indicated verbal language alone conveys enough clues to validly measure certain psychological constructs (Boyd et al., 2015; Tausczik & Pennebaker, 2010). Still, the dictionary method uses full transcripts as a 'bag of words', which is a simple text analysis method that ignores important factors like the order of the words. This method does not take the context into account (e.g., the words said before or after the distinct words). Further studies might apply more advanced machine learning models that would take the context and other features of behaviour into account.

Research findings showed that the prosody of teachers' language is crucial for students' educational and well-being outcomes (Rogerson & Dodd, 2005; Paulmann & Weinstein, 2022). Particularly, autonomy supportive communication has been characterised by using a quieter voice, slower speech rate, and less vocal energy. In contrast, controlling communication involves a higher energy, making the voice sound "harsher" (Weinstein et al., 2018). Also, research has shown that a controlling prosody is associated with lower basic psychological need satisfaction, well-being, and intention to disclose to teachers, whereas an autonomy supportive prosody is related to the satisfaction of autonomy and relatedness needs (Paulmann & Weinstein, 2022). While the dictionary is a simple and effective tool to assess motivational behaviour, it still remains an imperfect method of annotating behaviour. The dictionary method used in chapter 3 ignores teacher prosody, tone, and pitch, all of which likely influence student motivation. Future studies on teacher motivational behaviour may be able to align the ratings of the dictionary against verbal characteristics, like those from Weinstein et al. (2018). Doing so may be computationally feasible and efficient, while also integrating a wider range of important characteristics of teachers' behaviour.

In the systematic review, I showed that thousands of examples of each behaviour are required to train advanced, flexible machine learning methods. The review showed that these methods are often more accurate than simple methods, like dictionaries, but a few steps would be required to build such a system using the findings from this thesis. For example, researchers could now more easily create annotated examples of teacher behaviours using my classification. There are many existing studies with observational data from teacher behaviour (Lonsdale et al., 2013). Coding those data using my expert-derived classification would provide hundreds of examples of each behaviour in the classification. Then, instead of a dictionary, researchers might use those examples to train a more flexible machine learning model to classify those examples. For example, based on the findings of my systematic review, we might use those examples to train a support vector machine that classifies teacher sentences on the classification system, replicating methods used in psychotherapy (e.g., Tanana et al., 2016). If those models accurately classify behaviours from the transcripts, then the model may be put 'into production': it may be used to provide teachers with feedback about how to become more autonomy supportive, or could replace human coders in studies assessing teacher behaviour.

Conclusion

Accurate and rapid coding of teachers' motivational behaviour has been increasingly important to scale up interventions, to advance theoretical constructs, and to provide specific and individualised feedback. Through this thesis, I enabled this process by developing an automated method that would efficiently analyse teachers' motivational behaviour. To do so, I first conducted a systematic review (Chapter 2) of the applications of automated coding models to assess helping professionals' interpersonal interactions. The findings showed that automated methods were applied in psychotherapy, medical settings and a few in education. These methods showed promising results, and under some circumstances, achieved near-human performance. My dictionary study (Chapter 3) replicated some of these findings in education, where I was able to assess teacher motivational behaviour with observer-level agreement using an expert derived dictionary. However, my systematic review showed that the most sophisticated and accurate models usually applied a standardised behavioural coding measure. Thus, in Chapter 4, I created such a classification and laid the platform for future research to code discrete motivational behaviours. This classification of teacher motivational behaviours includes detailed descriptions, function descriptions, and examples for each behaviour. The classification may help pre-service and in-service teachers to learn need-supportive and thwarting behaviours. It may help the synthesis of the literature on the effectiveness of each motivational behaviour, and the replication and translation of effective interventions. Furthermore, both the dictionary and the classification may enable more reliable and valid feedback for teachers. Overall, these three studies help researchers and teachers to better understand adaptive motivational behaviour that supports the needs of students and creates environments where students are motivated and engaged in learning.

References

Aafjes-van Doorn, K., Kamsteeg, C., Bate, J., & Aafjes, M. (2021). A scoping review of machine learning in psychotherapy research. *Psychotherapy Research: Journal of the Society for Psychotherapy Research*, *31*(1), 92–116. https://doi.org/10.1080/10503307.2020.1808729

Abou Omar, K. B. (2018). XGBoost and LGBM for Porto Seguro's Kaggle challenge: A comparison. *Preprint Semester Project*. https://pub.tik.ee.ethz.ch/students/2017-HS/SA-2017-98.pdf

Abraham, C., & Michie, S. (2008). A taxonomy of behavior change techniques used in interventions. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, *27*(3), 379–387. https://doi.org/10.1037/0278-6133.27.3.379

Adamou, M., Antoniou, G., Greasidou, E., Lagani, V., Charonyktakis, P., & Tsamardinos, I. (2018). Mining Free-Text Medical Notes for Suicide Risk Assessment. *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, 1–8. https://doi.org/10.1145/3200947.3201020

Adie, J. W., Duda, J. L., & Ntoumanis, N. (2008). Autonomy support, basic need satisfaction and the optimal functioning of adult male and female sport participants: A test of basic needs theory. *Motivation and Emotion*, *32*(3), 189–199. https://doi.org/10.1007/s11031-008-9095-z

Aelterman, N., Vansteenkiste, M., Haerens, L., Soenens, B., Fontaine, J. R. J., & Reeve, J. (2019). Toward an integrative and fine-grained insight in motivating and demotivating teaching styles: The merits of a circumplex approach. *Journal of Educational Psychology*, *111*(3), 497–521. https://doi.org/10.1037/edu0000293

Aelterman, N., Vansteenkiste, M., Van Keer, H., & Haerens, L. (2016). Changing teachers' beliefs regarding autonomy support and structure: The role of experienced psychological need satisfaction in teacher training. *Psychology of Sport and Exercise*, *23*, 64–72. https://doi.org/10.1016/j.psychsport.2015.10.007

Ahmadi, A., Noetel, M., Schellekens, M., Parker, P., Antczak, D., Beauchamp, M., Dicke, T., Diezmann, C., Maeder, A., Ntoumanis, N., Yeung, A., & Lonsdale, C. (2021). A systematic review of machine learning for assessment and feedback of treatment fidelity. *Intervencion Psicosocial*. https://doi.org/10.5093/pi2021a4

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology*, *84*(3), 261. http://psycnet.apa.org/journals/edu/84/3/261.html?uid=1993-03487-001

Arnold, K. A., Turner, N., Barling, J., Kelloway, E. K., & McKee, M. C. (2007). Transformational leadership and psychological well-being: the mediating role of meaningful work. *Journal of Occupational Health Psychology*, *12*(3), 193–203. https://doi.org/10.1037/1076-8998.12.3.193

Ashford, S., Edmunds, J., & French, D. P. (2010). What is the best way to change self-efficacy to promote lifestyle and recreational physical activity? A systematic review with meta-analysis. *British Journal of Health Psychology*, *15*(Pt 2), 265–288. https://doi.org/10.1348/135910709X461752

Atkins, D., Rubin, N., Steyvers, M., Doeden, A., Baucom, R., & Christensen, A. (2012). Topic models: A novel method for modeling couple and family text data. *Journal of Family Psychology: JFP: Journal of the Division of Family Psychology of the American Psychological Association* , *26*(5), 816–827. https://doi.org/10.1037/a0029607

Atkins, D., Steyvers, M., Imel, Z., & Smyth, P. (2014). Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science: IS*, *9*, 49. https://doi.org/10.1186/1748-5908-9-49

Babineau, J. (2014). Product review: covidence (systematic review software). *Journal of the Canadian Health Libraries Association/Journal de l'Association Des Bibliothèques de La Santé Du Canada*, *35*(2), 68–71. https://journals.library.ualberta.ca/jchla/index.php/jchla/article/view/22892/17064

Bakeman, R., & Quera, V. (2011). *Sequential Analysis and Observational Methods for the Behavioral Sciences*. Cambridge University Press. https://play.google.com/store/books/details?id=_yOZAKivQfIC

Baker, J., Lovell, K., & Harris, N. (2006). How expert are the experts? An exploration of the concept of "expert"within Delphi panel techniques. *Nurse Researcher*, *14*(1). https://journals.rcni.com/doi/pdfplus/10.7748/nr2006.10.14.1.59.c6010

Bantum, E. O., & Owen, J. E. (2009). Evaluating the validity of computerized content analysis programs for identification of emotional expression in cancer narratives. *Psychological*

Assessment, *21*(1), 79–88. https://doi.org/10.1037/a0014643

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

Barth, J., Munder, T., Gerger, H., Nüesch, E., Trelle, S., Znoj, H., Jüni, P., & Cuijpers, P. (2013). Comparative efficacy of seven psychotherapeutic interventions for patients with depression: a network meta-analysis. *PLoS Medicine*, *10*(5), e1001454. https://doi.org/10.1371/journal.pmed.1001454

Bartholomew, K. J., Ntoumanis, N., Mouratidis, A., Katartzi, E., Thøgersen-Ntoumani, C., & Vlachopoulos, S. (2018). Beware of your teaching style: A school-year long investigation of controlling teaching and student motivational experiences. *Learning and Instruction*, *53*, 50–63. https://doi.org/10.1016/j.learninstruc.2017.07.006

Bartholomew, K. J., Ntoumanis, N., & Th∅gersen-Ntoumani, C. (2009). A review of controlling motivational strategies from a self-determination theory perspective: implications for sports coaches. *International Review of Sport and Exercise Psychology*, *2*(2), 215–233. https://doi.org/10.1080/17509840903235330

Bartholomew, K., Ntoumanis, N., Ryan, R., Bosch, J., & Thøgersen-Ntoumani, C. (2011). Self-determination theory and diminished functioning: the role of interpersonal control and psychological need thwarting. *Personality & Social Psychology Bulletin*, *37*(11), 1459–1473. https://doi.org/10.1177/0146167211413125

Bartholomew, K., Ntoumanis, N., Ryan, R., & Thøgersen-Ntoumani, C. (2011). Psychological Need Thwarting in the Sport Context: Assessing the Darker Side of Athletic Experience. *Journal of Sport and Exercise Psychology*, *33*(1), 75–102. https://doi.org/10.1123/jsep.33.1.75

Barwick, M. A., Bennett, L. M., Johnson, S. N., McGowan, J., & Moore, J. E. (2012). Training health and mental health professionals in motivational interviewing: A systematic review. *Children and*

Youth Services Review, *34*(9), 1786–1795. https://doi.org/10.1016/j.childyouth.2012.05.012

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, Articles*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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., 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/abs/10.1111/j.1758-0854.2011.01048.x

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., & 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

Belland, B. R., Walker, A. E., & Kim, N. J. (2017). A Bayesian Network Meta-Analysis to Synthesize the Influence of Contexts of Scaffolding Use on Cognitive Outcomes in STEM Education. *Review of Educational Research*, *87*(6), 1042–1081. https://doi.org/10.3102/0034654317723009

Bellg, A. J., Borrelli, B., Resnick, B., Hecht, J., Minicucci, D. S., Ory, M., Ogedegbe, G., Orwig, D., Ernst, D., Czajkowski, S., & Treatment Fidelity Workgroup of the NIH Behavior Change Consortium. (2004). Enhancing treatment fidelity in health behavior change studies: best practices and recommendations from the NIH Behavior Change Consortium. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, *23*(5), 443–451. https://doi.org/10.1037/0278-6133.23.5.443

Belmont, M., Skinner, E., Wellborn, J., & Connell, J. (1988). *Teacher as social context: A measure of student perceptions of teacher provision of involvement, structure, and autonomy support*. Tech.

rep.

Bhavsar, N., Ntoumanis, N., Quested, E., Gucciardi, D. F., Thøgersen-Ntoumani, C., Ryan, R. M., Reeve, J., Sarrazin, P., & Bartholomew, K. J. (2019). Conceptualizing and testing a new tripartite measure of coach interpersonal behaviors. *Psychology of Sport and Exercise*, *44*, 107–120. https://doi.org/10.1016/j.psychsport.2019.05.006

Biddle, S., Wang, C., Chatzisarantis, N., & Spray, C. (2003). Motivation for physical activity in young people: entity and incremental beliefs about athletic ability. *Journal of Sports Sciences*, *21*(12), 973–989. https://doi.org/10.1080/02640410310001641377

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

Blanchard, N., Donnelly, P. J., Olney, A. M., Samei, B., Ward, B., Sun, X., Kelly, S., Nystrand, M., & D'Mello, S. K. (2016a). Semi-Automatic Detection of Teacher Questions from Human-Transcripts of Audio in Live Classrooms. *Proceedings of the 9th International Conference on Educational Data Mining*. https://eric.ed.gov/?id=ED592742

Blanchard, N., Donnelly, P., Olney, A., Samei, B., Ward, B., Sun, X., Kelly, S., Nystrand, M., & D'Mello, S. K. (2016b). Identifying teacher questions using automatic speech recognition in classrooms. *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 191–201. https://www.aclweb.org/anthology/W16-3623.pdf

Blanck, P., Perleth, S., Heidenreich, T., Kröger, P., Ditzen, B., Bents, H., & Mander, J. (2018). Effects of mindfulness exercises as stand-alone intervention on symptoms of anxiety and depression: Systematic review and meta-analysis. *Behaviour Research and Therapy*, *102*, 25–35. https://doi.org/10.1016/j.brat.2017.12.002

Bogler, R. (2001). The Influence of Leadership Style on Teacher Job Satisfaction. *Educational Administration Quarterly: EAQ*, *37*(5), 662–683. https://doi.org/10.1177/00131610121969460

Borrelli, B. (2011). The Assessment, Monitoring, and Enhancement of Treatment Fidelity In Public Health Clinical Trials. *Journal of Public Health Dentistry*, *71*(s1), S52–S63. https://doi.org/10.1111/j.1752-7325.2011.00233.x

Bouchet-Valat, M. (2014). SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library. *R Package Version 0. 5*, *1*.

Boyce, B. A., Gano-Overway, L. A., & Campbell, A. L. (2009). Perceived Motivational Climate's Influence on Goal Orientations, Perceived Competence, and Practice Strategies across the Athletic Season. *Journal of Applied Sport Psychology*, *21*(4), 381–394. https://doi.org/10.1080/10413200903204887

Boyd. (2017). Psychological Text Analysis in the Digital Humanities. In S. Hai-Jew (Ed.), *Data Analytics in Digital Humanities* (pp. 161–189). Springer International Publishing. https://doi.org/10.1007/978-3-319-54499-1_7

Boyd, R., Wilson, S., Pennebaker, J., Kosinski, M., Stillwell, D., & Mihalcea, R. (2015). Values in Words: Using Language to Evaluate and Understand Personal Values. *Proceedings of the International AAAI Conference on Web and Social Media*, *9*(1), 31–40. https://ojs.aaai.org/index.php/ICWSM/article/view/14589

Bradshaw, E., Sahdra, B., Ciarrochi, J., Parker, P., Martos, T., & Ryan, R. (2021). A configural approach to aspirations: The social breadth of aspiration profiles predicts well-being over and above the intrinsic and extrinsic aspirations that comprise the profiles. *Journal of Personality and Social Psychology*, *120*(1), 226–256. https://doi.org/10.1037/pspp0000374

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Brown, B. (1968). *Delphi process: a methodology used for the elicitation of opinions of experts*. RAND Corporation. https://www.rand.org/content/dam/rand/pubs/papers/2006/P3925.pdf

Browne, M. W. (2000). Cross-Validation Methods. *Journal of Mathematical Psychology*, *44*(1), 108–132. https://doi.org/10.1006/jmps.1999.1279

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2005.14165

Bryan, C. L., & Solmon, M. A. (2007). Self-Determination in Physical Education: Designing Class Environments to Promote Active Lifestyles. *Journal of Teaching in Physical Education: JTPE*, *26*(3), 260–278. https://doi.org/10.1123/jtpe.26.3.260

Bureau, Howard, Chong, & Guay. (2022). Pathways to Student Motivation: A Meta-Analysis of Antecedents of Autonomous and Controlled Motivations. *Review of Educational Research*, *92*(1), 46–72. https://doi.org/10.3102/00346543211042426

Burnette, J. L., O'Boyle, E. H., VanEpps, E. M., Pollack, J. M., & Finkel, E. J. (2013). Mind-sets matter: a meta-analytic review of implicit theories and self-regulation. *Psychological Bulletin*, *139*(3), 655–701. https://doi.org/10.1037/a0029531

Burns, J. M. (1978). Leadership New York. *NY: Harper and Row Publishers*.

Can, Atkins, & Narayanan. (2015). A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations. *Sixteenth Annual Conference of the International Speech Communication Association*. https://188.166.204.102/archive/interspeech_2015/papers/i15_0339.pdf

Can, D., Marín, R., Georgiou, P., Imel, Z., Atkins, D., & Narayanan, S. (2016). "It sounds like...": A natural language processing approach to detecting counselor reflections in motivational interviewing. *Journal of Counseling Psychology*, *63*(3), 343–350. https://doi.org/10.1037/cou0000111

Can, Georgiou, P., Atkins, & Narayanan. (2012). *A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features*. *3*, 2251–2254. https://www.isca-speech.org/archive/interspeech_2012/i12_2254.html

Cao, Tanana, Imel, Poitras, Atkins, & Srikumar. (2019). Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5599–5611. https://doi.org/10.18653/v1/P19-1563

Carcone, A. I., Hasan, M., Alexander, G. L., Dong, M., Eggly, S., Brogan Hartlieb, K., Naar, S., MacDonell, K., & Kotov, A. (2019). Developing Machine Learning Models for Behavioral Coding. *Journal of Pediatric Psychology*, *44*(3), 289–299. https://doi.org/10.1093/jpepsy/jsy113

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research: JMLR*, *11*, 2079–2107. http://www.jmlr.org/papers/volume11/cawley10a/cawley10a.pdf

Chakravarthula, S. N., Gupta, R., Baucom, B., & Georgiou, P. (2015). A language-based generative model framework for behavioral analysis of couples' therapy. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2090–2094. https://doi.org/10.1109/ICASSP.2015.7178339

Chatzisarantis, N., & Hagger, M. (2009). Effects of an intervention based on self-determination theory on self-reported leisure-time physical activity participation. *Psychology & Health*, *24*(1), 29–48. https://doi.org/10.1080/08870440701809533

Chazan, D. J., Pelletier, G. N., & Daniels, L. M. (2022). Achievement Goal Theory Review: An Application to School Psychology. *Canadian Journal of School Psychology*, *37*(1), 40–56. https://doi.org/10.1177/08295735211058319

Chen, B., Vansteenkiste, M., Beyers, W., & Boone, L. (2015). Basic psychological need satisfaction, need frustration, and need strength across four cultures. *Motivation and Emotion*. https://idp.springer.com/authorize/casa?redirect_uri=https://link.springer.com/article/10.1007/s11031-014-9450-1&casa_token=diGsZiO-lqUAAAAA:OsdQOGqvX6YMa889DPrc7wCrhLdMjtln4ecag88oZKajIp2_LQPEH6y4u95x85QaVXedtw6RsCNU_08m

Chen, B., Vansteenkiste, M., Beyers, W., Boone, L., Deci, E., Van der Kaap-Deeder, J., Duriez, B., Lens, W., Matos, L., Mouratidis, A., Ryan, R., Sheldon, K., Soenens, B., Van Petegem, S., & Verstuyf, J. (2015). Basic psychological need satisfaction, need frustration, and need strength across four cultures. *Motivation and Emotion*, *39*(2), 216–236. https://doi.org/10.1007/s11031-014-9450-1

Cheon, S. H., Reeve, J., Lee, Y., Ntoumanis, N., Gillet, N., Kim, B. R., & Song, Y.-G. (2019). Expanding autonomy psychological need states from two (satisfaction, frustration) to three (dissatisfaction): A classroom-based intervention study. *Journal of Educational Psychology*, *111*(4), 685–702. https://doi.org/10.1037/edu0000306

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://www.ncbi.nlm.nih.gov/pubmed/22691399

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., … Fiedel, N. (2022). PaLM: Scaling Language Modeling with Pathways. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2204.02311

Christian, B. (2020). *The alignment problem: Machine learning and human values*. WW Norton & Company. https://unicenter.pt/sites/default/files/webform/cv/pdf-the-alignment-problem-machine-learning-and-human-values-brian-christian-pdf-download-free-book-e02f078.pdf

Collins, S. E., Chawla, N., Hsu, S. H., Grow, J., Otto, J. M., & Marlatt, G. A. (2009). Language-based measures of mindfulness: initial validity and clinical utility. *Psychology of Addictive Behaviors: Journal of the Society of Psychologists in Addictive Behaviors*, *23*(4), 743–749. https://doi.org/10.1037/a0017579

Costa, S., Soenens, B., Gugliandolo, M. C., Cuzzocrea, F., & Larcan, R. (2015). The Mediating Role of Experiences of Need Satisfaction in Associations Between Parental Psychological Control and Internalizing Problems: A Study Among Italian College Students. *Journal of Child and Family Studies*, *24*(4), 1106–1116. https://doi.org/10.1007/s10826-014-9919-2

Cox, A., & Williams, L. (2008). The roles of perceived teacher support, motivational climate, and psychological need satisfaction in students' physical education motivation. *Journal of Sport & Exercise Psychology*, *30*(2), 222–239. https://www.ncbi.nlm.nih.gov/pubmed/18490792

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., Petticrew, M., & Medical Research Council Guidance. (2008). Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ* , *337*, a1655. https://doi.org/10.1136/bmj.a1655

Creasey, G., & Jarvis, P. A. (2012). *Adolescent Development and School Achievement in Urban Communities: Resilience in the Neighborhood*. Routledge. https://market.android.com/details?id=book-kF56DoVwfeIC

Cross, K. P. (1977). Not can, but will college teaching be improved? *New Directions for Higher*

Education, *1977*(17), 1–15. https://doi.org/10.1002/he.36919771703

Czechowski, K., Miranda, D., & Sylvestre, J. (2016). Like a rolling stone: A mixed-methods approach to linguistic analysis of Bob Dylan's lyrics. *Psychology of Aesthetics, Creativity, and the Arts*, *10*(1), 99. http://psycnet.apa.org/journals/aca/10/1/99.html?uid=2015-55837-001

de Boer, H., Donker, A. S., & van der Werf, M. P. C. (2014). Effects of the attributes of educational interventions on students' academic performance. *Review of Educational Research*, *84*(4), 509–545. https://doi.org/10.3102/0034654314540006

Deci, E. L., & Ryan, R. M. (2002). Overview of self-determination theory: An organismic dialectical perspective. In E. Deci & R. Ryan (Eds.), *Handbook of self-determination research* (pp. 3–33). University of Rochester Press. https://books.google.com/books?hl=en&lr=&id=DcAe2b7L-RgC&oi=fnd&pg=PA3&dq=Overview+of+self-determination+theory+an+organismic+dialectical+perspective&ots=dszQ1KX6Zf&sig=jD2XGC4EPXT-ASTZuY_QJRmuYIA

Deci, E., & Ryan, R. (1985). The general causality orientations scale: Self-determination in personality. *Journal of Research in Personality*, *19*(2), 109–134. https://doi.org/10.1016/0092-6566(85)90023-6

Deci, E., & Ryan, R. (2000). The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*(4), 227–268. https://www.tandfonline.com/doi/abs/10.1207/s15327965pli1104_01

Deci, E., & Ryan, R. (2002). Overview of self-determination theory: An organismic dialectical perspective. *Handbook of Self-Determination Research*, 3–33. https://books.google.com.au/books?hl=en&lr=&id=DcAe2b7L-RgC&oi=fnd&pg=PA3&dq=Ryan,+R.+M.,+%26+Deci,+E.+L.+(2002).+Overview+of+self-determination+theory:+An+organismic+dialectical+perspective.+In+E.+L.+Deci+%26+R.+M.+Ryan+(Eds.),+Handbook+of+self-determination+research+(pp.+3-33).+Rochester,+NY:+University+of+Rochester+Press.&ots=dryQ2BX61l&sig=6a1mlL2iRNxm6qeT6o3n2l2GiPA

Deci, & Ryan. (1985). *Intrinsic Motivation and Self-Determination in Human Behavior*. Springer

Science & Business Media. https://market.android.com/details?id=book-p96Wmn-ER4QC

Delbecq, A. L., Van de Ven, A. H., & Gustafson, D. H. (1975). *Group techniques for program planning: A guide to nominal group and Delphi processes*. Scott, Foresman,. http://eduq.info/xmlui/handle/11515/11368

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

Diamond, I. R., Grant, R. C., Feldman, B. M., Pencharz, P. B., Ling, S. C., Moore, A. M., & Wales, P. W. (2014). Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies. *Journal of Clinical Epidemiology*, *67*(4), 401–409. https://doi.org/10.1016/j.jclinepi.2013.12.002

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

Donnelly, P. J., Blanchard, N., Olney, A. M., Kelly, S., Nystrand, M., & D'Mello, S. K. (2017). Words matter: automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, 218–227. https://doi.org/10.1145/3027385.3027417

Donohoe, H. M., & Needham, R. D. (2009). Moving best practice forward: Delphi characteristics, advantages, potential problems, and solutions. *International Journal of Tourism Research*, *11*(5), 415–437. https://doi.org/10.1002/jtr.709

Duda, J. L., Papaioannou, A. G., Appleton, P. R., Quested, E., and Krommidas, C. (2014). "Creating adaptive motivational climates in sport and physical education," in *Routledge Companion to Sport and Exercise Psychology: Global Perspectives and Fundamental Concepts*, eds A. G. Papaioannou and D. Hackfort (New York, NY: Routledge), 544–558.

Duncan, E., Nicol, M. M., & Ager, A. (2004). Factors that constitute a good cognitive behavioural treatment manual: A Delphi study. *Behavioural and Cognitive Psychotherapy*. http://dspace.stir.ac.uk/handle/1893/11937

Dweck, C. (1999). *Self-Theories: Their role in motivation, personality and development. 1999*. Hove:

Psychology Press.

Dweck, C., & Leggett, L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, *95*(2), 256. http://psycnet.apa.org/journals/rev/95/2/256.html?uid=1988-29536-001

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., Jha, S., Agrawal, M., Dziurzynski, L. A., Sap, M., Weeg, C., Larson, E. E., Ungar, L. H., & Seligman, M. E. P. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, *26*(2), 159–169. https://doi.org/10.1177/0956797614557867

Escartí, A., & Gutiérrez, M. (2001). Influence of the motivational climate in physical education on the intention to practice physical activity or sport. *European Journal of Sport Science: EJSS: Official Journal of the European College of Sport Science*, *1*(4), 1–12. https://doi.org/10.1080/17461390100071406

Explosion. (2022). *Prodigy · An annotation tool for AI, Machine Learning & NLP*. Prodigy. https://prodi.gy/

Fairburn, C. G., & Cooper, Z. (2011). Therapist competence, therapy quality, and therapist training. *Behaviour Research and Therapy*, *49*(6-7), 373–378. https://doi.org/10.1016/j.brat.2011.03.005

Fairburn, C. G., & Patel, V. (2017). The impact of digital technology on psychological treatments and their dissemination. *Behaviour Research and Therapy*, *88*, 19–25. https://doi.org/10.1016/j.brat.2016.08.012

Falk, J. (2013). *We will rock you : A diachronic corpus-based analysis of linguistic features in rock lyrics* [diva-portal.org]. http://www.diva-portal.org/smash/record.jsf?pid=diva2:605003

Fast, E., Chen, B., & Bernstein, M. S. (2017). Lexicons on demand: neural word embeddings for large-scale text analysis. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 4836–4840. https://www.ijcai.org/proceedings/2017/0677.pdf

Feinerer, I., Hornik, K., Wallace, M., & Hornik, M. K. (2020). *Package “wordnet.”* http://cran.fhcrc.org/web/packages/wordnet/wordnet.pdf

Fisher, R. J. (1993). Social Desirability Bias and the Validity of Indirect Questioning. *The Journal of Consumer Research*, *20*(2), 303–315. https://doi.org/10.1086/209351

Flemotomos, N., Martinez, V. R., Gibson, J., Atkins, D. C., Creed, T., & Narayanan, S. S. (2018). Language Features for Automated Evaluation of Cognitive Behavior Psychotherapy Sessions. *INTERSPEECH*, 1908–1912. 10.21437/Interspeech.2018-1518

Franco, E., & Coterón, J. (2017). The Effects of a Physical Education Intervention to Support the Satisfaction of Basic Psychological Needs on the Motivation and Intentions to be Physically Active. *Journal of Human Kinetics*, *59*, 5–15. https://doi.org/10.1515/hukin-2017-0143

French, D. P., Olander, E. K., Chisholm, A., & Mc Sharry, J. (2014). Which behaviour change techniques are most effective at increasing older adults' self-efficacy and physical activity behaviour? A systematic review. *Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine*, *48*(2), 225–234. https://doi.org/10.1007/s12160-014-9593-z

Froiland, J. M., & Worrell, F. C. (2016). Intrinsic motivation, learning goals, engagement, and achievement in a diverse high school. *Psychology in the Schools*, *53*(3), 321–336. https://doi.org/10.1002/pits.21901

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

Gallo, C., Pantin, H., Villamar, J., Prado, G., Tapia, M., Ogihara, M., Cruden, G., & Brown, C. (2015). Blending Qualitative and Computational Linguistics Methods for Fidelity Assessment: Experience with the Familias Unidas Preventive Intervention. *Administration and Policy in Mental Health*, *42*(5), 574–585. https://doi.org/10.1007/s10488-014-0538-4

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

García, S., Luengo, J., & Herrera, F. (2014). *Data Preprocessing in Data Mining*. Springer International Publishing. https://doi.org/10.1007/978-3-319-10247-4

Gaut, Steyvers, Imel, Atkins, & Smyth. (2017). Content Coding of Psychotherapy Transcripts Using Labeled Topic Models. *IEEE Journal of Biomedical and Health Informatics*, *21*(2), 476–487.

https://doi.org/10.1109/JBHI.2015.2503985

Gibson, Can, Georgiou, Atkins, & Narayanan. (2017). Attention Networks for Modeling Behaviors in Addiction Counseling. *Interspeech 2017*, 3251–3255. https://doi.org/10.21437/Interspeech.2017-218

Gibson, Can, Xiao, Imel, Atkins, Georgiou, & Narayanan. (2016). A Deep Learning Approach to Modeling Empathy in Addiction Counseling. *Interspeech 2016*, *2016*, 1447–1451. https://doi.org/10.21437/Interspeech.2016-554

Gillet, N., Vallerand, R. J., & Lafrenière, M.-A. K. (2012). Intrinsic and extrinsic school motivation as a function of age: the mediating role of autonomy support. *Social Psychology of Education: An International Journal*, *15*(1), 77–95. https://doi.org/10.1007/s11218-011-9170-2

Gnambs, T., & Hanfstingl, B. (2016). The decline of academic motivation during adolescence: An accelerated longitudinal cohort analysis on the effect of psychological need satisfaction. *Educational Psychology Review*, *36*(9), 1691–1705. https://doi.org/10.1080/01443410.2015.1113236

Goldberg, Flemotomos, & Martinez. (2020). Machine learning and natural language processing in psychotherapy research: Alliance as example use case. *Journal of Counseling*. https://psycnet.apa.org/record/2020-46941-003

Goldberg, Rousmaniere, Miller, Whipple, Nielsen, Hoyt, & Wampold. (2016). Do psychotherapists improve with time and experience? A longitudinal analysis of outcomes in a clinical setting. *Journal of Counseling Psychology*, *63*(1), 1–11. https://doi.org/10.1037/cou0000131

Golin, C. E., Liu, H., Hays, R. D., Miller, L. G., Beck, C. K., Ickovics, J., Kaplan, A. H., & Wenger, N. S. (2002). A prospective study of predictors of adherence to combination antiretroviral medication. *Journal of General Internal Medicine*, *17*(10), 756–765. https://doi.org/10.1046/j.1525-1497.2002.11214.x

González-Cutre, D., Romero-Elías, M., Jiménez-Loaisa, A., Beltrán-Carrillo, V. J., & Hagger, M. S. (2020). Testing the need for novelty as a candidate need in basic psychological needs theory. *Motivation and Emotion*, *44*(2), 295–314. https://doi.org/10.1007/s11031-019-09812-7

Gottfried, A. E., Gottfried, A. W., Morris, P. E., & Cook, C. R. (2008). Low academic intrinsic motivation as a risk factor for adverse educational outcomes: A longitudinal study from early childhood through early adulthood. *Academic Motivation and the Culture of School in Childhood and Adolescence.*, *320*, 36–69. https://doi.org/10.1093/acprof:oso/9780195326819.003.0003

Goudas, M., Biddle, S., & Fox, K. (1994). Perceived locus of causality, goal orientations, and perceived competence in school physical education classes. *The British Journal of Educational Psychology*, *64 (Pt 3)*, 453–463. https://www.ncbi.nlm.nih.gov/pubmed/7811633

Graf, E.-M., Sator, M., & Spranz-Fogasy, T. (2014). *Discourses of Helping Professions*. John Benjamins Publishing Company. https://play.google.com/store/books/details?id=hKK2BQAAQBAJ

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. https://doi.org/10.1037/a0015141

Greenhalgh, T., & Peacock, R. (2005). Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. *BMJ* , *331*(7524), 1064–1065. https://doi.org/10.1136/bmj.38636.593461.68

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Grouzet, F., Kasser, T., Ahuvia, A., Dols, J., Kim, Y., Lau, S., Ryan, R., Saunders, S., Schmuck, P., & Sheldon, K. (2005). The structure of goal contents across 15 cultures. *Journal of Personality and Social Psychology*, *89*(5), 800–816. https://doi.org/10.1037/0022-3514.89.5.800

Guay, F. (2022). Applying Self-Determination Theory to Education: Regulations Types, Psychological Needs, and Autonomy Supporting Behaviors. *Canadian Journal of School Psychology*, *37*(1), 75–92. https://doi.org/10.1177/08295735211055355

Gunnell, K. E., Crocker, P. R. E., Mack, D. E., Wilson, P. M., & Zumbo, B. D. (2014). Goal contents, motivation, psychological need satisfaction, well-being and physical activity: A test of self-determination theory over 6 months. *Psychology of Sport and Exercise*, *15*(1), 19–29.

https://doi.org/10.1016/j.psychsport.2013.08.005

Haerens, L., Aelterman, N., Van den Berghe, L., De Meyer, J., Soenens, B., & Vansteenkiste, M. (2013a). Observing physical education teachers’ need-supportive interactions in classroom settings. *Journal of Sport & Exercise Psychology*, *35*(1), 3–17. https://doi.org/10.1123/jsep.35.1.3

Haerens, L., Aelterman, N., Van den Berghe, L., De Meyer, J., Soenens, B., & Vansteenkiste, M. (2013b). Observing physical education teachers’ need-supportive interactions in classroom settings. *Journal of Sport & Exercise Psychology*, *35*(1), 3–17. https://doi.org/10.1123/jsep.35.1.3

Haerens, L., Aelterman, N., Vansteenkiste, M., Soenens, B., & Van Petegem, S. (2015). Do perceived autonomy-supportive and controlling teaching relate to physical education students’ motivational experiences through unique pathways? Distinguishing between the bright and dark side of motivation. *Psychology of Sport and Exercise*, *16*, 26–36. https://www.sciencedirect.com/science/article/pii/S1469029214001204

Haerens, L., Vansteenkiste, M., De Meester, A., Delrue, J., Tallir, I., Vande Broek, G., Goris, W., & Aelterman, N. (2018). Different combinations of perceived autonomy support and control: identifying the most optimal motivating style. *Physical Education and Sport Pedagogy*, *23*(1), 16–36. https://doi.org/10.1080/17408989.2017.1346070

Hagger, M., Cameron, L., Hamilton, K., Hankonen, N., & Lintunen, T. (2020). *The Handbook of Behavior Change*. Cambridge University Press. https://play.google.com/store/books/details?id=IfEFEAAAQBAJ

Hagger, M., Chatzisarantis, N., Culverhouse, T., & Biddle, S. (2003). The processes by which perceived autonomy support in physical education promotes leisure-time physical activity intentions and behavior: a trans-contextual model. *Journal of Educational Psychology*, *95*(4), 784. http://doi.apa.org/journals/edu/95/4/784.html

Hagger, M., Chatzisarantis, N., & Harris, J. (2006). From psychological need satisfaction to intentional behavior: Testing a motivational sequence in two behavioral contexts. *Personality & Social Psychology Bulletin*, *32*(2), 131–148. https://doi.org/10.1177/0146167205279905

Hagger, M., Chatzisarantis, N. L. D., Hein, V., Soós, I., Karsai, I., Lintunen, T., & Leemans, S. (2009). Teacher, peer and parent autonomy support in physical education and leisure-time physical activity: A trans-contextual model of motivation in four nations. *Psychology & Health*, *24*(6), 689–711. https://doi.org/10.1080/08870440801956192

Hagger, M., Koch, S., & Chatzisarantis, N. (2015). The effect of causality orientations and positive competence-enhancing feedback on intrinsic motivation: A test of additive and interactive effects. *Personality and Individual Differences*, *72*, 107–111. https://doi.org/10.1016/j.paid.2014.08.012

Hagger, M., & Weed, M. (2019). DEBATE: Do interventions based on behavioral theory work in the real world? *The International Journal of Behavioral Nutrition and Physical Activity*, *16*(1), 36. https://doi.org/10.1186/s12966-019-0795-4

Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., Lee, M. J., & Asadi, H. (2019). Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods. *AJR. American Journal of Roentgenology*, *212*(1), 38–43. https://doi.org/10.2214/AJR.18.20224

Hardcastle, S., Fortier, M., Blake, N., & Hagger, M. (2017). Identifying content-based and relational techniques to change behaviour in motivational interviewing. *Health Psychology Review*, *11*(1), 1–16. https://doi.org/10.1080/17437199.2016.1190659

Hasan, M., Kotov, A., Carcone, A., Dong, M., Naar, S., & Hartlieb, K. B. (2016). A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories. *Journal of Biomedical Informatics*, *62*, 21–31. https://doi.org/10.1016/j.jbi.2016.05.004

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. New York, NY: Springer. http://thuvien.thanglong.edu.vn:8081/dspace/handle/DHTL_123456789/4053

Hattie, J. (2008). *Visible Learning*. https://doi.org/10.4324/9780203887332

Hausner, E., Guddat, C., Hermanns, T., Lampert, U., & Waffenschmidt, S. (2016). Prospective comparison of search strategies for systematic reviews: an objective approach yielded higher sensitivity than a conceptual one. *Journal of Clinical Epidemiology*, *77*, 118–124. https://doi.org/10.1016/j.jclinepi.2016.05.002

Hein, V., Müür, M., & Koka, A. (2004). Intention to be Physically Active after School Graduation and Its Relationship to Three Types of Intrinsic Motivation. *European Physical Education Review*, *10*(1), 5–19. https://doi.org/10.1177/1356336X04040618

Henderlong, J., & Lepper, M. R. (2002). The effects of praise on children's intrinsic motivation: a review and synthesis. *Psychological Bulletin*, *128*(5), 774–795. https://www.ncbi.nlm.nih.gov/pubmed/12206194

Hernandez-Orallo, J. (2012). *A unified view of performance metrics: Translating threshold choice into expected classification loss*. jmlr.org. https://www.jmlr.org/papers/volume13/hernandez-orallo12a/hernandez-orallo12a.pdf

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2021). *Cochrane Handbook for Systematic Reviews of Interventions* (Vol. 6.2). Cochrane. www.training.cochrane.org/handbook

Hinde, S., & Spackman, E. (2015). Bidirectional citation searching to completion: An exploration of literature searching methods. *PharmacoEconomics*, *33*(1), 5–11. https://doi.org/10.1007/s40273-014-0205-3

Hirsch, Soma, Merced, Kuo, Dembe, Caperton, Atkins, & Imel. (2018). " It's hard to argue with a computer" Investigating Psychotherapists' Attitudes towards Automated Evaluation. *Proceedings of the 2018 Designing Interactive Systems Conference*, 559–571. https://dl.acm.org/doi/abs/10.1145/3196709.3196776

Howard, J., Bureau, J, Guay, F., Chong, J., & Ryan, R. (2021). Student Motivation and Associated Outcomes: A Meta-Analysis From Self-Determination Theory. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *16*(6), 1300–1323. https://doi.org/10.1177/1745691620966789

Howard, M., Agarwal, G., & Hilts, L. (2009). Patient satisfaction with access in two interprofessional academic family medicine clinics. *Family Practice*, *26*(5), 407–412. https://doi.org/10.1093/fampra/cmp049

Howes, C., Purver, M., & McCabe, R. (2013). Using conversation topics for predicting therapy outcomes in schizophrenia. *Biomedical Informatics Insights*, *6*(Suppl 1), 39–50. https://doi.org/10.4137/BII.S11661

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

Huyghebaert-Zouaghi, T., Ntoumanis, N., Berjot, S., & Gillet, N. (2021). Advancing the Conceptualization and Measurement of Psychological Need States: A 3 × 3 Model Based on Self-Determination Theory. *Journal of Career Assessment*, *29*(3), 396–421. https://doi.org/10.1177/1069072720978792

Iliev, R., Dehghani, M., & Sagi, E. (2015). Automated text analysis in psychology: Methods, applications, and future developments. *Language and Cognition*, *7*(2), 265–290. https://www.cambridge.org/core/journals/language-and-cognition/article/automated-text-analysis-in-psychology-methods-applications-and-future-developments/97F60083638BD8E956958084BE1E049A

Imel, Pace, Soma, Tanana, Hirsch, Gibson, Georgiou, Narayanan, & Atkins. (2019). Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. *Psychotherapy* , *56*(2), 318–328. https://doi.org/10.1037/pst0000221

Imel, Z., Steyvers, M., & Atkins, D. (2015). Computational psychotherapy research: scaling up the evaluation of patient-provider interactions. *Psychotherapy* , *52*(1), 19–30. https://doi.org/10.1037/a0036841

Ivers, N., Jamtvedt, G., Flottorp, S., Young, J. M., Odgaard-Jensen, J., French, S. D., O'Brien, M. A., Johansen, M., Grimshaw, J., & Oxman, A. D. (2012). Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database of Systematic Reviews* , *6*, CD000259. https://doi.org/10.1002/14651858.CD000259.pub3

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

Jang, H., Kim, E. J., & Reeve, J. (2016). Why students become more engaged or more disengaged during the semester: A self-determination theory dual-process model. *Learning and Instruction*, *43*, 27–38. https://doi.org/10.1016/j.learninstruc.2016.01.002

Jang, H., Reeve, J., & Deci, E. (2010). Engaging students in learning activities: It is not autonomy support or structure but autonomy support and structure. *Journal of Educational Psychology*, *102*(3), 588. http://psycnet.apa.org/record/2010-15712-005

Jorm, A. F. (2015). Using the Delphi expert consensus method in mental health research. *The Australian and New Zealand Journal of Psychiatry*, *49*(10), 887–897. https://doi.org/10.1177/0004867415600891

Judge, T. A., & Piccolo, R. F. (2004). Transformational and transactional leadership: a meta-analytic test of their relative validity. *The Journal of Applied Psychology*, *89*(5), 755–768. https://doi.org/10.1037/0021-9010.89.5.755

Jurka, T. P., Collingwood, L., Boydstun, A. E., Grossman, E., & van Atteveldt, W. (2012). RTextTools: Automatic text classification via supervised learning. *R Package Version*, *1*(9).

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *The American Psychologist*, *64*(6), 515–526. https://doi.org/10.1037/a0016755

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: a flaw in human judgment*. Little, Brown. https://www.academia.edu/download/52417053/Kahneman_Noise.pdf

Kaplan, A., & Maehr, M. L. (2007). The Contributions and Prospects of Goal Orientation Theory. *Educational Psychology Review*, *19*(2), 141–184. https://doi.org/10.1007/s10648-006-9012-5

Kasser, T., & Ryan, R. (1996). Further Examining the American Dream: Differential Correlates of Intrinsic and Extrinsic Goals. *Personality & Social Psychology Bulletin*, *22*(3), 280–287. https://doi.org/10.1177/0146167296223006

Kazantzis, N. (2003). Therapist Competence in Cognitive-behavioural Therapies: Review of the Contemporary Empirical Evidence. In *Behaviour Change* (Vol. 20, Issue 1, pp. 1–12). https://doi.org/10.1375/bech.20.1.1.24845

Kazdin, A. E. (2017). Addressing the treatment gap: A key challenge for extending evidence-based psychosocial interventions. *Behaviour Research and Therapy*, *88*, 7–18. https://doi.org/10.1016/j.brat.2016.06.004

Keeney, S., Hasson, F., & McKenna, H. (2006). Consulting the oracle: Ten lessons from using the Delphi technique in nursing research. *Journal of Advanced Nursing*, *53*(2), 205–212. https://doi.org/10.1111/j.1365-2648.2006.03716.x

Keeney, S., Hasson, F., & McKenna, H. (2017). *The Delphi technique in nursing and health research*. http://dlib.sbmu.ac.ir/site/catalogue/166209

Keeney, S., Hasson, F., & McKenna, H. P. (2001). A critical review of the Delphi technique as a research methodology for nursing. *International Journal of Nursing Studies*, *38*(2), 195–200. https://doi.org/10.1016/S0020-7489(00)00044-4

Kopcha, T. J., & Sullivan, H. (2007). Self-presentation bias in surveys of teachers' educational technology practices. *Educational Technology Research and Development: ETR & D*, *55*(6), 627–646. https://doi.org/10.1007/s11423-006-9011-8

Kornhaber, R., Walsh, K., Duff, J., & Walker, K. (2016). Enhancing adult therapeutic interpersonal relationships in the acute health care setting: an integrative review. *Journal of Multidisciplinary Healthcare*, *9*, 537–546. https://doi.org/10.2147/JMDH.S116957

Korpershoek, H., Harms, T., de Boer, H., van Kuijk, M., & Doolaard, S. (2016). A meta-analysis of the effects of classroom management strategies and classroom management programs on students academic, behavioral, emotional, and motivational outcomes. *Review of Educational Research*, *86*(3), 643–680. https://doi.org/10.3102/0034654315626799

Kotov, A., Idalski Carcone, A., Dong, M., Naar-King, S., & Brogan, K. E. (2014). Towards automatic coding of interview transcripts for public health research. *Proceedings of the Big Data Analytic Technology For Bioinformatics and Heath Informatics Workshop (KDD-BHI) in Conjunction with ACM SIGKDD Conference on Knowledge Discovery and Data Mining, New York, NY.*

Krijgsman, C., Mainhard, T., Borghouts, L., van Tartwijk, J., & Haerens, L. (2020). Do goal clarification and process feedback positively affect students' need-based experiences? A quasi-experimental study grounded in self-determination theory. *Physical Education and Sport Pedagogy*, 1–21. https://doi.org/10.1080/17408989.2020.1823956

La Guardia, J., Ryan, R., Couchman, C., & Deci, E. (2000). Within-person variation in security of attachment: a self-determination theory perspective on attachment, need fulfillment, and well-being. *Journal of Personality and Social Psychology*, *79*(3), 367–384. https://www.ncbi.nlm.nih.gov/pubmed/10981840

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, *33*(2), 363–374. https://www.ncbi.nlm.nih.gov/pubmed/884196

Latvala, E., Vuokila-Oikkonen, P., & Janhonen, S. (2000). Videotaped recording as a method of participant observation in psychiatric nursing research. *Journal of Advanced Nursing*, *31*(5), 1252–1257. https://www.ncbi.nlm.nih.gov/pubmed/10840260

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). Big data. The parable of Google Flu: traps in big data analysis. *Science*, *343*(6176), 1203–1205. https://doi.org/10.1126/science.1248506

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

Legault, L. (2017). Self-Determination Theory. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of Personality and Individual Differences* (pp. 1–9). Springer International Publishing. https://doi.org/10.1007/978-3-319-28099-8_1162-1

Lepper, M. R., Corpus, J. H., & Iyengar, S. S. (2005). Intrinsic and extrinsic motivational orientations in the classroom: Age differences and academic correlates. *Journal of Educational Psychology*, *97*(2), 184–196. https://doi.org/10.1037/0022-0663.97.2.184

Link, L. J. (2022). *What Teachers Really Want When It Comes to Feedback*. ASCD. https://www.ascd.org/el/articles/what-teachers-really-want-when-it-comes-to-feedback

Liu, B. (2010). Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing*, *2*, 627–666. ftp://nozdr.ru/biblio/kolxo3/Cs/CsNl/Indurkhya%20N.,%20Damerau%20F.J.%20(eds.)%20Handbook%20of%20natural%20language%20processing%20(2ed.,%20CRC,%202010)(ISBN%209781420085921)(O)(692s)_CsNl_.pdf#page=653

Liukkonen, J., Barkoukis, V., Watt, A., and Jaakkola, T. (2010). Motivational climate and students' emotional experiences and effort in physical education. *J. Educ. Res.* 103, 295–308. doi: 10.1080/00220670903383044

Lonsdale, C., Rosenkranz, R. R., Peralta, L. R., Bennie, A., Fahey, P., & Lubans, D. R. (2013). A systematic review and meta-analysis of interventions designed to increase moderate-to-vigorous physical activity in school physical education lessons. *Preventive medicine*, 56(2), 152-161. doi: 10.1016/j.ypmed.2012.12.004

Loughlin, K. G., & Moore, L. F. (1979). Using Delphi to achieve congruent objectives and activities in a pediatrics department. *Journal of Medical Education*, *54*(2), 101–106. https://doi.org/10.1097/00001888-197902000-00006

Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, *5*(1), 96–116. https://doi.org/10.1002/lio2.354

Lum, K. (2017). *Limitations of mitigating judicial bias with machine learning Machine-learning algorithms trained with data that encode human bias will reproduce, not eliminate, the bias, says Kristian Lum*. NATURE PUBLISHING GROUP 75 VARICK ST, 9TH FLR, NEW YORK, NY 10013-1917 USA.

Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., Shilton, A., Yearwood, J., Dimitrova, N., Ho, T. B., Venkatesh, S., & Berk, M. (2016). Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *Journal of Medical Internet Research*, *18*(12), e323. https://doi.org/10.2196/jmir.5870

Macaskill, P., Gatsonis, C., Deeks, J., Harbord, R., & Takwoingi, Y. (2010). Cochrane handbook for systematic reviews of diagnostic test accuracy. *Version 0. 9. 0. London: The Cochrane Collaboration*. http://methods.cochrane.org/sites/methods.cochrane.org.sdt/files/public/uploads/Chapter%2010%20-%20Version%201.0.pdf

Madson, M. B., Loignon, A. C., & Lane, C. (2009). Training in motivational interviewing: a systematic review. *Journal of Substance Abuse Treatment*, *36*(1), 101–109. https://doi.org/10.1016/j.jsat.2008.05.005

Mageau, G. A., Ranger, F., Joussemet, M., Koestner, R., Moreau, E., & Forest, J. (2015). Validation of the Perceived Parental Autonomy Support Scale (P-PASS). *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement*, *47*(3), 251. http://psycnet.apa.org/record/2015-30412-004

Mahoney, J. W., Ntoumanis, N., Gucciardi, D. F., Mallett, C. J., & Stebbings, J. (2016). Implementing an Autonomy-Supportive Intervention to Develop Mental Toughness in Adolescent Rowers. *Journal of Applied Sport Psychology*, *28*(2), 199–215. https://doi.org/10.1080/10413200.2015.1101030

Marsh, H. W., Papaioannou, A., & Theodorakis, Y. (2006). Causal ordering of physical self-concept and exercise behavior: reciprocal effects model and the influence of physical education teachers. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, *25*(3), 316–328. https://doi.org/10.1037/0278-6133.25.3.316

Martin, A. J., Nejad, H. G., Colmar, S., & Liem, G. A. D. (2013). Adaptability: How students' responses to uncertainty and novelty predict their academic and non-academic outcomes. *Journal of Educational Psychology*, *105*(3), 728. http://psycnet.apa.org/record/2013-14504-001

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

McGlinchey, J. B., & Dobson, K. S. (2003). Treatment integrity concerns in cognitive therapy for depression. *Journal of Cognitive Psychotherapy*, *17*(4), 299–318. https://doi.org/10.1891/jcop.17.4.299.52543

Michie, S., Abraham, C., Whittington, C., McAteer, J., & Gupta, S. (2009). Effective techniques in healthy eating and physical activity interventions: A meta-regression. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, *28*(6), 690–701. https://doi.org/10.1037/a0016136

Michie, S., Ashford, S., Sniehotta, F. F., Dombrowski, S. U., Bishop, A., & French, D. P. (2011). A refined taxonomy of behaviour change techniques to help people change their physical activity and healthy eating behaviours: The CALO-RE taxonomy. *Psychology & Health*, *26*(11), 1479–1498. https://doi.org/10.1080/08870446.2010.540664

Michie, S., Carey, R. N., Johnston, M., Rothman, A. J., De Bruin, M., Kelly, M. P., & Connell, L. E. (2018). From theory-inspired to theory-based interventions: A protocol for developing and testing a methodology for linking behaviour change techniques to theoretical mechanisms of action. *Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine*, *52*(6), 501–512. https://academic.oup.com/abm/article-abstract/52/6/501/4737217

Michie, S., Fixsen, D., Grimshaw, J. M., & Eccles, M. P. (2009). Specifying and reporting complex behaviour change interventions: The need for a scientific method. *Implementation Science: IS*, *4*, 40. https://doi.org/10.1186/1748-5908-4-40

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., Eccles, M. P., Cane, J., & Wood, C. E. (2013). The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine*, *46*(1), 81–95. https://doi.org/10.1007/s12160-013-9486-6

Michie, S., Wood, C. E., Johnston, M., Abraham, C., Francis, J., & Hardeman, W. (2015). Behaviour change techniques: The development and evaluation of a taxonomic method for reporting and describing behaviour change interventions (a suite of five studies involving consensus methods, randomised controlled trials and analysis of qualitative data). *Health Technology Assessment*, *19*(99). https://openaccess.city.ac.uk/id/eprint/13435/

Miller, Moyers, Ernst, & Amrhein. (2003). Manual for the motivational interviewing skill code (MISC). *Unpublished Manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*. https://casaa.unm.edu/download/misc.pdf

Miller, W., & Rollnick, S. (1991). Motivational Interviewing: Preparing People to Change Addictive Behaviour, New York. *Guilford Press. Mueser, K. , Bellack, AS & Blanchard, JJ (1992). Comorbidity of Schizophrenia and Substance Abuse: Implications for Treatment. Journal of Consulting and Clinical Psychology*, *60*, 845–856.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word--emotion association lexicon.

Computational Intelligence. An International Journal, *29*(3), 436–465. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8640.2012.00460.x

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis: An Annual Publication of the Methodology Section of the American Political Science Association*, *16*(4), 372–403. https://doi.org/10.1093/pan/mpn018

Moore, C. M. (1987). Group techniques for idea building. *Applied Social Research Methods Series, Vol. 9.*, *143*. https://psycnet.apa.org/fulltext/1987-97921-000.pdf

Moreau, D., & Gamble, B. (2020). Conducting a meta-analysis in the age of open science: Tools, tips, and practical recommendations. *Psychological Methods*. https://doi.org/10.1037/met0000351

Morgan, K., Sproule, J., Weigand, D., & Carpenter, P. (2005). A computer-based observational assessment of the teaching behaviours that influence motivational climate in Physical Education. *Physical Education and Sport Pedagogy*, *10*(1), 83–105. https://doi.org/10.1080/1740898042000334926

Morton, K. L., Keith, S. E., & Beauchamp, M. R. (2010). Transformational teaching and physical activity: a new paradigm for adolescent health promotion? *Journal of Health Psychology*, *15*(2), 248–257. https://doi.org/10.1177/1359105309347586

Moyers, Miller, & Hendrickson. (2005). How does motivational interviewing work? Therapist interpersonal skill predicts client involvement within motivational interviewing sessions. *Journal of Consulting and Clinical Psychology*, *73*(4), 590–598. https://doi.org/10.1037/0022-006X.73.4.590

Moyers, Rowell, Manuel, Ernst, & Houck. (2016). The Motivational Interviewing Treatment Integrity Code (MITI 4): Rationale, Preliminary Reliability and Validity. *Journal of Substance Abuse Treatment*, *65*, 36–42. https://doi.org/10.1016/j.jsat.2016.01.001

Moyers, T., Martin, T., Manuel, J., Hendrickson, S., & Miller, W. (2005). Assessing competence in the use of motivational interviewing. *Journal of Substance Abuse Treatment*, *28*(1), 19–26. https://doi.org/10.1016/j.jsat.2004.11.001

Muijs, D. (2006). Measuring teacher effectiveness: Some methodological reflections. *Educational Research and Evaluation*, *12*(1), 53–74. https://doi.org/10.1080/13803610500392236

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) Psychology: Measuring and Mapping Scales of Cultural and Psychological Distance. *Psychological Science*, *31*(6), 678–701. https://doi.org/10.1177/0956797620916782

Nelson, L. K., Burk, D., Knudsen, M., & McCall, L. (2018). The Future of Coding: A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods. *Sociological Methods & Research*, 0049124118769114. https://doi.org/10.1177/0049124118769114

Newton, P. M., Da Silva, A., & Berry, S. (2020). The Case for Pragmatic Evidence-Based Higher Education: A Useful Way Forward? *Frontiers in Education*, *5*. https://doi.org/10.3389/feduc.2020.583157

Ng, A. (2019). Machine learning yearning: Technical strategy for ai engineers in the era of deep learning. *Retrieved Online at Https://www.mlyearning.org*.

Ng, J., Ntoumanis, N., Thøgersen-Ntoumani, C., Deci, E., Ryan, R., Duda, J., & Williams, G. (2012). Self-Determination Theory applied to health contexts: A meta-analysis. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *7*(4), 325–340. https://doi.org/10.1177/1745691612447309

Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological Review*, *91*(3), 328. http://psycnet.apa.org/record/1984-28719-001

Nicholls, J. G. (1989). *The Competitive Ethos and Democratic Education*. Harvard University Press. https://market.android.com/details?id=book-CmdUo6P9CL0C

Nicolas, G., Bai, X., & Fiske, S. (2019). Automated Dictionary Creation for Analyzing Text: An illustration from stereotype content. In *PsyArXiv*. https://doi.org/10.31234/osf.io/afm8k

Niemiec, C., & Ryan, R. (2009). Autonomy, competence, and relatedness in the classroom: Applying self-determination theory to educational practice. *Educational Research and Evaluation: An International Journal on Theory and Practice*, *7*(2), 133–144. https://doi.org/10.1177/1477878509104318

Niemiec, C., Ryan, R., & Deci, E. (2009). The Path Taken: Consequences of Attaining Intrinsic and Extrinsic Aspirations in Post-College Life. *Journal of Research in Personality*, *73*(3), 291–306. https://doi.org/10.1016/j.jrp.2008.09.001

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. (2001). A self-determination approach to the understanding of motivation in physical education. *The British Journal of Educational Psychology*, *71*(Pt 2), 225–242. https://www.ncbi.nlm.nih.gov/pubmed/11449934

Ntoumanis, N. (2002). Motivational clusters in a sample of British physical education classes. *Psychology of Sport and Exercise*, *3*(3), 177–194. https://doi.org/10.1016/S1469-0292(01)00020-6

Ntoumanis, N., Ng, J. Y. Y., Prestwich, A., Quested, E., Hancox, J. E., Thøgersen-Ntoumani, C., Deci, E. L., Ryan, R. M., Lonsdale, C., & Williams, G. C. (2020). A meta-analysis of self-determination theory-informed intervention studies in the health domain: Effects on motivation, health behavior, physical, and psychological health. *Health Psychology Review*, 1–31. https://www.tandfonline.com/doi/abs/10.1080/17437199.2020.1718529

Ntoumanis, N., & Standage, M. (2009). Motivation in physical education classes: A self-determination theory perspective. *School Field*, *7*(2), 194–202. https://doi.org/10.1177/1477878509104324

Ntoumanis, N., Thøgersen-Ntoumani, C., Quested, E., & Hancox, J. (2017). The effects of training group exercise class instructors to adopt a motivationally adaptive communication style. *Scandinavian Journal of Medicine & Science in Sports*, *27*(9), 1026–1034. https://doi.org/10.1111/sms.12713

Nystrand, M., Wu, L. L., Gamoran, A., Zeiser, S., & Long, D. A. (2003). Questions in Time:

Investigating the Structure and Dynamics of Unfolding Classroom Discourse. *Discourse Processes*, *35*(2), 135–198. https://doi.org/10.1207/S15326950DP3502_3

Olafsen, A., Deci, E., & Halvari, H. (2018). Basic psychological needs and work motivation: A longitudinal test of directionality. *Motivation and Emotion*, *42*(2), 178–189. https://doi.org/10.1007/s11031-017-9646-2

Ommundsen, Y. (2001). Pupilsí Affective Responses in Physical Education Classes: the Association of Implicit Theories of the Nature of Ability and Achievement Goals. *European Physical Education Review*, *7*(3), 219–242. https://doi.org/10.1177/1356336X010073001

Öst, L.-G., & Ollendick, T. H. (2017). Brief, intensive and concentrated cognitive behavioral treatments for anxiety disorders in children: A systematic review and meta-analysis. *Behaviour Research and Therapy*, *97*, 134–145. https://doi.org/10.1016/j.brat.2017.07.008

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.

Park, J., Kotzias, D., Kuo, P., Logan, R. L., Iv, Merced, K., Singh, S., Tanana, M., Karra Taniskidou, E., Lafata, J. E., Atkins, D. C., Tai-Seale, M., Imel, Z. E., & Smyth, P. (2019). Detecting conversation topics in primary care office visits from transcripts of patient-provider interactions. *Journal of the American Medical Informatics Association: JAMIA*, *26*(12), 1493–1504. https://doi.org/10.1093/jamia/ocz140

Parsons, J. T., Rosof, E., Punzalan, J. C., & Di Maria, L. (2005). Integration of motivational interviewing and cognitive behavioral therapy to improve HIV medication adherence and reduce substance use among HIV-positive men and women: results of a pilot project. *AIDS Patient Care and STDs*, *19*(1), 31–39. https://doi.org/10.1089/apc.2005.19.31

Patall, E. A., Steingut, R. R., Vasquez, A. C., Trimble, S. S., Pituch, K. A., & Freeman, J. L. (2018). Daily autonomy supporting or thwarting and students' motivation and engagement in the high school science classroom. *Journal of Educational Psychology*, *110*(2), 269–288. https://doi.org/10.1037/edu0000214

Patall, E. A., Vasquez, A. C., Steingut, R. R., Trimble, S. S., & Pituch, K. A. (2017). Supporting and thwarting autonomy in the high school science classroom. *Cognition and Instruction*, *35*(4), 337–362. https://doi.org/10.1080/07370008.2017.1358722

Patridge, & Bardyn. (2018). Research Electronic Data Capture (REDCap). *Journal of the Medical Library Association: JMLA*, *106*(1), 142. https://doi.org/10.5195/jmla.2018.319

Paulmann, S., & Weinstein, N. (2022). Teachers' motivational prosody: A pre-registered experimental test of children's reactions to tone of voice used by teachers. *The British journal of educational psychology*, 10.1111/bjep.12567. Advance online publication. https://doi.org/10.1111/bjep.12567

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. repositories.lib.utexas.edu. https://repositories.lib.utexas.edu/handle/2152/31333

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001. http://downloads.liwc.net.s3.amazonaws.com/LIWC2015_OperatorManual.pdf

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language. use: our words, our selves. *Annual Review of Psychology*, *54*, 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Perepletchikova, F., Hilt, L. M., Chereji, E., & Kazdin, A. E. (2009). Barriers to implementing treatment integrity procedures: survey of treatment outcome researchers. *Journal of Consulting and Clinical Psychology*, *77*(2), 212–218. https://doi.org/10.1037/a0015232

Perepletchikova, F., & Kazdin, A. E. (2005). Treatment Integrity and Therapeutic Change: Issues and Research Recommendations. *Clinical Psychology: Science and Practice*, *12*(4), 365–383. https://doi.org/10.1093/clipsy.bpi045

Perepletchikova, F., Treat, T. A., & Kazdin, A. E. (2007). Treatment integrity in psychotherapy research: analysis of the studies and examination of the associated factors. *Journal of Consulting and Clinical Psychology*, *75*(6), 829–841. https://doi.org/10.1037/0022-006X.75.6.829

Pérez-Rosas, Sun, Li, Wang, Resnicow, & Mihalcea. (2019). *Analyzing the quality of counseling conversations: The tell-tale signs of high-quality counseling*. 3742–3748.

https://www.scopus.com/inward/record.uri?eid=2-s2.0-85059886164&partnerID=40&md5=dac1c5b64d2f82e022ecb542d991b9db

Pérez-Rosas, V., Mihalcea, R., Resnicow, K., Singh, S., An, L., Goggin, K. J., & Catley, D. (2017). Predicting Counselor Behaviors in Motivational Interviewing Encounters. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1128–1137. https://www.aclweb.org/anthology/E17-1106

Perlich, C. (2009). *Learning Curves in Machine Learning* (No. RC24756). IBM Research Division. https://domino.research.ibm.com/library/cyberdig.nsf/papers/491B767CE4518A4585257576006BCD2D/$File/rc24756.pdf

Petterson, T. J. (2008). *Word Usage and Thematic Content of Song Lyric Analyses*. https://diginole.lib.fsu.edu/islandora/object/fsu%3A176435/

Piccolo, R. F., & Colquitt, J. A. (2006). Transformational Leadership and Job Behaviors: The Mediating Role of Core Job Characteristics. *Academy of Management Journal*, *49*(2), 327–340. https://doi.org/10.5465/amj.2006.20786079

Pietraszkiewicz, A., Formanowicz, M., Gustafsson Sendén, M., Boyd, R. L., Sikström, S., & Sczesny, S. (2019). The big two dictionaries: Capturing agency and communion in natural language. *European Journal of Social Psychology*, *49*(5), 871–887. https://doi.org/10.1002/ejsp.2561

Plucker, J. A., & Makel, M. C. (2021). Replication is important for educational psychology: Recent developments and key issues. *Educational Psychologist*, *56*(2), 90–100. https://doi.org/10.1080/00461520.2021.1895796

Powell, C. (2003). The Delphi technique: Myths and realities. *Journal of Advanced Nursing*, *41*(4), 376–382. https://doi.org/10.1046/j.1365-2648.2003.02537.x

Prestwich, A., Sniehotta, F. F., Whittington, C., Dombrowski, S. U., Rogers, L., & Michie, S. (2014). Does theory influence the effectiveness of health behavior interventions? Meta-analysis. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, *33*(5), 465–474. https://doi.org/10.1037/a0032853

Prowse, P.-T. D., & Nagel, T. (2015). A Meta-Evaluation: The Role of Treatment Fidelity within Psychosocial Interventions during the Last Decade. *Journal of Psychiatry*, *18*(2).

https://doi.org/10.4172/psychiatry.1000251

Prowse, P.-T. D., Nagel, T., Meadows, G. N., & Enticott, J. C. (2015). Treatment Fidelity Over the Last Decade in Psychosocial Clinical Trials Outcome Studies: A Systematic Review. *Journal of Psychiatry*, *18*(2). https://doi.org/10.4172/psychiatry.1000258

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. *Proceedings of the First Instructional Conference on Machine Learning*, *242*, 29–48. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.1424&rep=rep1&type=pdf

Reeve, J. (2009). Why teachers adopt a controlling motivating style toward students and how they can become more autonomy supportive. *Educational Psychologist*, *44*(3), 159–175. https://doi.org/10.1080/00461520903028990

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/10.1007/978-1-4614-2018-7_7

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

Reeve, J., & Halusic, M. (2009). How K-12 teachers can put self-determination theory principles into practice. *Educational Research and Evaluation: An International Journal on Theory and Practice*, *7*(2), 145–154. https://doi.org/10.1177/1477878509104319

Reeve, J., & Jang, H. (2006). What teachers say and do to support students' autonomy during a learning activity. *Journal of Educational Psychology*, *98*(1), 209–218. https://doi.org/10.1037/0022-0663.98.1.209

Reeve, J., Jang, H., Carrell, D., Jeon, S., & Barch, J. (2004). Enhancing students' engagement by increasing teachers' autonomy support. *Motivation and Emotion*, *28*(2), 147–169. https://doi.org/10.1023/B:MOEM.0000032312.95499.6f

Reeve, J., & Sickenius, B. (1994). Development and Validation of a Brief Measure of the Three Psychological Needs Underlying Intrinsic Motivation: The Afs Scales. *Educational and Psychological Measurement*, *54*(2), 506–515. https://doi.org/10.1177/0013164494054002025

Reiser, R. P., & Milne, D. L. (2014). A systematic review and reformulation of outcome evaluation in clinical supervision: Applying the fidelity framework. *Training and Education in Professional Psychology*, *8*(3), 149–157. https://doi.org/10.1037/tep0000031

Riniolo, T. C., Johnson, K. C., Sherman, T. R., & Misso, J. A. (2006). Hot or not: do professors perceived as physically attractive receive higher student evaluations? *The Journal of General Psychology*, *133*(1), 19–35. https://doi.org/10.3200/GENP.133.1.19-35

Robbins, M. L., Mehl, M. R., Smith, H. L., & Weihs, K. L. (2013). Linguistic indicators of patient, couple, and family adjustment following breast cancer. *Psycho-Oncology*, *22*(7), 1501–1508. https://doi.org/10.1002/pon.3161

Roberts, G. C. (2001). Understanding the dynamics of motivation in physical activity: The influence of achievement goals on motivational processes. *Advances in Motivation in Sport and Exercise*, 1–50.

Robins, R. W., & Pals, J. L. (2002). Implicit Self-Theories in the Academic Domain: Implications for Goal Orientation, Attributions, Affect, and Self-Esteem Change. *Self and Identity: The Journal of the International Society for Self and Identity*, *1*(4), 313–336. https://doi.org/10.1080/15298860290106805

Rodriguez-Quintana, N., & Lewis, C. C. (2018). Observational Coding Training Methods for CBT Treatment Fidelity: A Systematic Review. *Cognitive Therapy and Research*, *42*(4), 358–368. https://doi.org/10.1007/s10608-018-9898-5

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

Ross, J. A., & Gray, P. (2006). Transformational leadership and teacher commitment to organizational values: The mediating effects of collective teacher efficacy. *School Effectiveness and School Improvement*, *17*(2), 179–199. https://doi.org/10.1080/09243450600565795

Rude, S., Gortner, E.-M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, *18*(8), 1121–1133. https://doi.org/10.1080/02699930441000030

Russell, S., & Norvig, P. (2002). *Artificial intelligence: a modern approach*. https://research.google/pubs/pub27702.pdf

Ryan, P., Luz, S., Albert, P., Vogel, C., Normand, C., & Elwyn, G. (2019). Using artificial intelligence to assess clinicians' communication skills. *BMJ* , *364*, l161. https://doi.org/10.1136/bmj.l161

Ryan, R. (1991). A motivational approach to self: Integration in personality edward l., deci and. *Perspectives on Motivation*, *38*(237), 237–288. https://books.google.com.au/books?hl=en&lr=&id=veZlIWFOSGgC&oi=fnd&pg=PA237&dq=Deci,+E.+L.,+%26+Ryan,+R.+M.+(1991).+A+motivational+approach+to+self:+Integration+in+personality.+In+R.+Dienstibier+(Ed.),+Nebraska+Symposium+on+Motivation:+Vol.+38.+Perspectives+on+motivation+(pp.+237-288).+Lincoln,+NE:+University+of+Nebraska+Press.&ots=Ir8uU5cATY&sig=wJrgHdBrTaX2JFbzBxXqGnFy3-o

Ryan, R., & Deci, E. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *The American Psychologist*, *55*(1), 68–78. https://www.ncbi.nlm.nih.gov/pubmed/11392867

Ryan, R., & Deci, E. (2017). *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness*. Guilford Publications. https://play.google.com/store/books/details?id=GF0ODQAAQBAJ

Ryan, R., & Deci, E. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational*

Psychology, *61*, 101860. https://doi.org/10.1016/j.cedpsych.2020.101860

Saleh, M., Lazonder, A. W., & De Jong, T. (2005). Effects of within-class ability grouping on social interaction, achievement, and motivation. *Instructional Science*, *33*(2), 105–119. https://doi.org/10.1007/s11251-004-6405-z

Samei, B., Olney, A. M., Kelly, S., Nystrand, M., & D'Mello, S. (2014). Domain Independent Assessment of Dialogic Properties of Classroom Discourse. *Grantee Submission*. https://eric.ed.gov/?id=ED566380

Sarrazin, P. G., Tessier, D. P., Pelletier, L. G., Trouilloud, D. O., & Chanal, J. P. (2006). The effects of teachers' expectations about students' motivation on teachers' autonomy-supportive and controlling behaviors. *International Journal of Sport and Exercise Psychology*, *4*(3), 283–301. https://www.tandfonline.com/doi/abs/10.1080/1612197X.2006.9671799

Schwalbe, C. S., Oh, H. Y., & Zweben, A. (2014). Sustaining motivational interviewing: a meta-analysis of training studies. *Addiction* , *109*(8), 1287–1294. https://doi.org/10.1111/add.12558

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine*, *49*(9), 1426–1448. https://doi.org/10.1017/S0033291719000151

Sheldon, K. M. (2011). Integrating behavioral-motive and experiential-requirement perspectives on psychological needs: a two process model. *Psychological Review*, *118*(4), 552–569. https://doi.org/10.1037/a0024758

Sheldon, K. M., Abad, N., & Omoile, J. (2009). Testing Self-Determination Theory via Nigerian and Indian adolescents. *International Journal of Behavioral Development*, *33*(5), 451–459. https://doi.org/10.1177/0165025409340095

Sheldon, K. M., & Filak, V. (2008). Manipulating autonomy, competence, and relatedness support in a game-learning context: New evidence that all three needs matter. *British Journal of Social Psychology*, *47*(2), 267–283. https://onlinelibrary.wiley.com/doi/abs/10.1348/014466607X238797

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Silge, J., & Robinson, D. (2017). *Text Mining with R: A Tidy Approach*. O'Reilly. https://play.google.com/store/books/details?id=7bQzMQAACAAJ

Simon, P. (2013). *Too Big to Ignore: The Business Case for Big Data*. John Wiley & Sons. https://market.android.com/details?id=book-zAemCgAAQBAJ

Simonsen, B., Fairbanks, S., Briesch, A., Myers, D., & Sugai, G. (2008). Evidence-based Practices in Classroom Management: Considerations for Research to Practice. *Education & Treatment of Children*, *31*(3), 351–380. http://www.jstor.org/stable/42899983

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

Skipp, A., & Tanner, E. (2015). The Visible Classroom: Evaluation Report and Executive Summary. *Education Endowment Foundation*. https://eric.ed.gov/?id=ED581106

Slemp, G., Kern, M., Patrick, K., & Ryan, R. (2018). Leader autonomy support in the workplace: A meta-analytic review. *Motivation and Emotion*. https://doi.org/10.1007/s11031-018-9698-y

Slemp, G. R., Field, J. G., & Cho, A. S. H. (2020). A meta-analysis of autonomous and controlled forms of teacher motivation. *Journal of Vocational Behavior*, *121*, 103459. https://doi.org/10.1016/j.jvb.2020.103459

Smale-Jacobse, A. E., Meijer, A., Helms-Lorenz, M., & Maulana, R. (2019). Differentiated instruction in secondary education: A systematic review of research evidence. *Frontiers in Psychology*, *10*, 2366. https://doi.org/10.3389/fpsyg.2019.02366

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

Smith, N., Tessier, D., Tzioumakis, Y., Quested, E., Appleton, P., Sarrazin, P., Papaioannou, A., & Duda, J. L. (2015). Development and validation of the multidimensional motivational climate observation system. *Journal of Sport & Exercise Psychology*, *37*(1), 4–22.

https://doi.org/10.1123/jsep.2014-0059

Standage, M., Duda, J. L., & Ntoumanis, N. (2005). A test of self-determination theory in school physical education. *The British Journal of Educational Psychology*, *75*(3), 411–433. https://onlinelibrary.wiley.com/doi/abs/10.1348/000709904X22359

Standage, M., Duda, J. L., & Ntoumanis, N. (2006). Students' motivational processes and their relationship to teacher ratings in school physical education: a self-determination theory approach. *Research Quarterly for Exercise and Sport*, *77*(1), 100–110. https://doi.org/10.1080/02701367.2006.10599336

Standage, M., Gillison, F., & Treasure, D. C. (2007). *PsycNET*. psycnet.apa.org. http://psycnet.apa.org/record/2007-05407-005

Standage, M., & Treasure, D. C. (2002). Relationship among achievement goal orientations and multidimensional situational motivation in physical education. *The British Journal of Educational Psychology*, *72*(Pt 1), 87–103. https://www.ncbi.nlm.nih.gov/pubmed/11916466

Stewart, J., O'Halloran, C., Harrigan, P., Spencer, J. A., Barton, J. R., & Singleton, S. J. (1999). Identifying appropriate tasks for the preregistration year: modified Delphi technique. *BMJ* , *319*(7204), 224–229. https://doi.org/10.1136/bmj.319.7204.224

Street, R. L., Jr, Makoul, G., Arora, N. K., & Epstein, R. M. (2009). How does communication heal? Pathways linking clinician-patient communication to health outcomes. *Patient Education and Counseling*, *74*(3), 295–301. https://doi.org/10.1016/j.pec.2008.11.015

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

Sylvester, B. D., Curran, T., Standage, M., Sabiston, C. M., & Beauchamp, M. R. (2018). Predicting exercise motivation and exercise behavior: A moderated mediation model testing the interaction between perceived exercise variety and basic psychological needs satisfaction. *Psychology of Sport and Exercise*, *36*, 50–56. https://www.sciencedirect.com/science/article/pii/S1469029217304843?casa_token=smcHlZt_WUYAAAAA:-f5w2DrxPHbIgRnz94hFVYXao7N0ywTMnYU-

3JuhVHnxNBD3b5vYqedZB1tBUcRn-f3CHKgH8SwA

Tai, A. M. Y., Albuquerque, A., Carmona, N. E., Subramanieapillai, M., Cha, D. S., Sheko, M., Lee, Y., Mansur, R., & McIntyre, R. S. (2019). Machine learning and big data: Implications for disease modeling and therapeutic discovery in psychiatry. *Artificial Intelligence in Medicine*, *99*, 101704. https://doi.org/10.1016/j.artmed.2019.101704

Tanana, M. (2021, November 3). *Predicting CBT fidelity like a human*. Lyssn | Intelligent Counselling Recording Platform. https://www.lyssn.io/blog/predicting-cbt-fidelity-like-a-human

Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., & Srikumar, V. (2016). A Comparison of Natural Language Processing Methods for Automated Coding of Motivational Interviewing. *Journal of Substance Abuse Treatment*, *65*, 43–50. https://doi.org/10.1016/j.jsat.2016.01.006

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *The Journal of the International Association of Medical Science Educators: JIAMSE*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Taylor, G., Jungert, T., Mageau, G. A., Schattke, K., Dedic, H., Rosenfield, S., & Koestner, R. (2014). A self-determination theory approach to predicting school achievement over time: The unique role of intrinsic motivation. *Contemporary Educational Psychology*, *39*(4), 342–358. https://doi.org/10.1016/j.cedpsych.2014.08.002

Teixeira, P. J., Marques, M. M., Silva, M. N., Brunet, J., Duda, J., Haerens, L., La Guardia, J., Lindwall, M., Londsdale, C., Markland, D., Michie, S., Moller, A. C., Ntoumanis, N., Patrick, H., Reeve, J., Ryan, R. M., Sebire, S., Standage, M., Vansteenkinste, M., … Hagger, M. S. (2020). Classification of techniques used in self-determination theory-based interventions in health contexts: An expert consensus study. *Motivation Science*. https://doi.org/10.1037/mot0000172

Teixeira, P. J., Marques, M. M., Silva, M. N., Brunet, J., Duda, J., Haerens, L., La Guardia, J., Lindwall, M., Lonsdale, C., Markland, D., & Others. (2020). A classification of motivation and behavior change techniques used in self-determination theory-based interventions in health contexts. *Motivation Science*. https://purehost.bath.ac.uk/ws/files/203332754/SDT_BCTs_2020.pdf

Tessier, D., Sarrazin, P., & Ntoumanis, N. (2010a). The effect of an intervention to improve newly qualified teachers' interpersonal style, students motivation and psychological need satisfaction in sport-based physical education. *Contemporary Educational Psychology*, *35*(4), 242–253. https://www.sciencedirect.com/science/article/pii/S0361476X10000305

Tessier, D., Sarrazin, P., & Ntoumanis, N. (2010b). The effect of an intervention to improve newly qualified teachers' interpersonal style, students motivation and psychological need satisfaction in sport-based physical education. *Contemporary Educational Psychology*, *35*(4), 242–253. https://doi.org/10.1016/j.cedpsych.2010.05.005

Tracey, T. J. G., Wampold, B. E., Lichtenberg, J. W., & Goodyear, R. K. (2014). Expertise in psychotherapy: an elusive goal? *The American Psychologist*, *69*(3), 218–229. https://doi.org/10.1037/a0035099

Treasure, D. C., & Robert, G. C. (2001). Students' Perceptions of the Motivational Climate, Achievement Beliefs, and Satisfaction in Physical Education. *Research Quarterly for Exercise and Sport*, *72*(2), 165–175. https://doi.org/10.1080/02701367.2001.10608946

Trevelyan, E. G., & Robinson, N. (2015). Delphi methodology in health research: how to do it? *European Journal of Integrative Medicine*, *7*(4), 423–428. https://www.sciencedirect.com/science/article/pii/S1876382015300160?casa_token=tEWd6PXIVoIAAAAA:avglxzj2y6unwZN46QHM6i3kegXTGCHp7XWzzx6Wx8JOOvhNeH5qm6WFpOCUKarVaAyblKys

United Nations. (2015). *Transforming our world: the 2030 Agenda for Sustainable Development*. United Nations. https://sdgs.un.org/2030agenda

Van den Berghe, L., Cardon, G., Tallir, I., Kirk, D., & Haerens, L. (2016). Dynamics of need-supportive and need-thwarting teaching behavior: the bidirectional relationship with student engagement and disengagement in the beginning of a lesson. *Physical Education and Sport Pedagogy*, *21*(6), 653–670. https://doi.org/10.1080/17408989.2015.1115008

Van den Berghe, L., Soenens, B., Vansteenkiste, M., Aelterman, N., Cardon, G., Tallir, I. B., & Haerens, L. (2013). Observed need-supportive and need-thwarting teaching behavior in physical education: Do teachers' motivational orientations matter? *Psychology of Sport and Exercise*, *14*(5), 650–661. https://doi.org/10.1016/j.psychsport.2013.04.006

Vansteenkiste, M., & Deci, E. (2003). Competitively Contingent Rewards and Intrinsic Motivation: Can Losers Remain Motivated? *Motivation and Emotion*, *27*(4), 273–299. https://doi.org/10.1023/A:1026259005264

Vansteenkiste, M., Lens, W., & Deci, E. (2006). Intrinsic Versus Extrinsic Goal Contents in Self-Determination Theory: Another Look at the Quality of Academic Motivation. *Educational Psychologist*, *41*(1), 19–31. https://doi.org/10.1207/s15326985ep4101_4

Vansteenkiste, M., Niemiec, C., & Soenens, B. (2010). The development of the five mini-theories of self-determination theory: an historical overview, emerging trends, and future directions. In *The Decade Ahead: Theoretical Perspectives on Motivation and Achievement* (pp. 105–165). emeraldinsight.com. https://doi.org/10.1108/S0749-7423(2010)000016A007

Vansteenkiste, M., & Ryan, R. (2013). On psychological growth and vulnerability: Basic psychological need satisfaction and need frustration as a unifying principle. *Journal of Psychotherapy Integration*, *23*(3), 263. http://psycnet.apa.org/fulltext/2013-20985-001.html

Vansteenkiste, M., Ryan, R., & Soenens, B. (2020). Basic psychological need theory: Advancements, critical themes, and future directions. *Motivation and Emotion*, *44*(1), 1–31. https://doi.org/10.1007/s11031-019-09818-1

Vansteenkiste, M., Simons, J., Lens, W., Soenens, B., Matos, L., & Lacante, M. (2004). Less is sometimes more: Goal content matters. *Journal of Educational Psychology*, *96*(4), 755. http://psycnet.apa.org/record/2004-21454-014

Vansteenkiste, M., Timmermans, T., Lens, W., Soenens, B., & Van den Broeck, A. (2008). Does extrinsic goal framing enhance extrinsic goal-oriented individuals' learning and performance? An experimental test of the match perspective versus self-determination theory. *Journal of Educational Psychology*, *100*(2), 387. http://psycnet.apa.org/record/2008-05694-009

Vasalou, A., Gill, A. J., Mazanderani, F., Papoutsi, C., & Joinson, A. (2011). Privacy dictionary: A new resource for the automated content analysis of privacy. *Journal of the American Society for Information Science. American Society for Information Science*, *62*(11), 2095–2105. https://doi.org/10.1002/asi.21610

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. (2020). Self-determination theory applied to physical education: A systematic review and meta-analysis. *Journal of Educational Psychology*, *112*(7), 1444–1469. https://doi.org/10.1037/edu0000420

Vlachopoulos, S. P., Asci, F. H., Cid, L., Ersoz, G., González-Cutre, D., Moreno-Murcia, J. A., & Moutão, J. (2013). Cross-cultural invariance of the basic psychological needs in exercise scale and need satisfaction latent mean differences among Greek, Spanish, Portuguese and Turkish samples. *Psychology of Sport and Exercise*, *14*(5), 622–631. https://doi.org/10.1016/j.psychsport.2013.03.002

Wallace, B. C., Laws, M. B., Small, K., Wilson, I. B., & Trikalinos, T. A. (2014). Automatically annotating topics in transcripts of patient-provider interactions via machine learning. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, *34*(4), 503–512. https://doi.org/10.1177/0272989X13514777

Waller, G. (2009). Evidence-based treatment and therapist drift. *Behaviour Research and Therapy*, *47*(2), 119–127. https://doi.org/10.1016/j.brat.2008.10.018

Waller, G., & Turner, H. (2016). Therapist drift redux: Why well-meaning clinicians fail to deliver evidence-based therapy, and how to get back on track. *Behaviour Research and Therapy*, *77*, 129–137. https://doi.org/10.1016/j.brat.2015.12.005

Wang, G., Oh, I., Courtright, S., & Colbert, A. (2011). Transformational Leadership and Performance Across Criteria and Levels: A Meta-Analytic Review of 25 Years of Research. *Group & Organization Management*, *36*(2), 223–270. https://doi.org/10.1177/1059601111401017

Wang, W., Hernandez, I., Newman, D., He, J., & Bian, J. (2016). Twitter analysis: Studying US weekly trends in work stress and emotion. *Applied Psychology = Psychologie Appliquee*, *65*(2), 355–378. https://iaap-journals.onlinelibrary.wiley.com/doi/abs/10.1111/apps.12065

Wang, Z., Pan, X., Miller, K., & Cortina, K. (2014). Automatic classification of activities in classroom discourse. *Computers & Education*, *78*, 115–123. https://doi.org/10.1016/j.compedu.2014.05.010

Warburton, V. E., & Spray, C. M. (2017). Implicit Theories of Ability in Physical Education: Current Issues and Future Directions. *Journal of Teaching in Physical Education: JTPE*, *36*(3), 252–261. https://doi.org/10.1123/jtpe.2017-0043

Ward, L., Stebbings, S., Sherman, K. J., Cherkin, D., & Baxter, G. D. (2014). Establishing key components of yoga interventions for musculoskeletal conditions: a Delphi survey. *BMC Complementary and Alternative Medicine*, *14*, 196. https://doi.org/10.1186/1472-6882-14-196

Weinstein, N., Przybylski, A., & Ryan, R. (2012). The index of autonomous functioning: Development of a scale of human autonomy. *Journal of Research in Personality*, *46*(4), 397–413. https://doi.org/10.1016/j.jrp.2012.03.007

Weinstein, N., Zougkou, K., & Paulmann, S. (2018). You 'have' to hear this: Using tone of voice to motivate others. Journal of experimental psychology. *Human perception and performance*, *44*(6), 898–913. https://doi.org/10.1037/xhp0000502

Wendler, C., Glazer, N., & Cline, F. (2019). Examining the calibration process for raters of the GRE ® general test. *ETS Research Report Series*, *2019*(1), 1–19. https://doi.org/10.1002/ets2.12245

Williams, G., & Deci, E. (1996). Internalization of biopsychosocial values by medical students: a test of self-determination theory. *Journal of Personality and Social Psychology*, *70*(4), 767–779. https://doi.org/10.1037//0022-3514.70.4.767

Wilson, P. M., Todd Rogers, W., Rodgers, W. M., & Cameron Wild, T. (2006). The Psychological Need Satisfaction in Exercise Scale. *Journal of Sport and Exercise Psychology*, *28*(3), 231–251. https://doi.org/10.1123/jsep.28.3.231

Wisniewski, B., Zierer, K., & Hattie, J. (2019). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Xiao, Huang, Imel, Atkins, Georgiou, & Narayanan. (2016). A technology prototype system for rating therapist empathy from audio recordings in addiction counseling. *PeerJ. Computer Science*, *2*(4). https://doi.org/10.7717/peerj-cs.59

Xiao, Imel, Georgiou, Atkins, & Narayanan. (2015a). “ Rate my therapist”: automated detection of empathy in drug and alcohol counseling via speech and language processing. *PloS One*, *10*(12), e0143055. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0143055

Xiao, Imel, Georgiou, Atkins, & Narayanan. (2015b). “Rate My Therapist”: Automated Detection of Empathy in Drug and Alcohol Counseling via Speech and Language Processing. *PloS One*, *10*(12), e0143055. https://doi.org/10.1371/journal.pone.0143055

Yarkoni, T. (2010). Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, *44*(3), 363–373. https://doi.org/10.1016/j.jrp.2010.04.001

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

Research Portfolio Appendix

Systematic Review (Chapter 2)

The systematic review in Chapter 2 has been published in Psychosocial Intervention.

The published copy is available here:

Ahmadi, A., Noetel, M., Schellekens, M., Parker, P., Antczak, D., Beauchamp, M., Dicke, T., Diezmann, C., Maeder, A., Ntoumanis, N., Yeung, A., and Lonsdale, C. (2021). A Systematic Review of Machine Learning for Assessment and Feedback of Treatment Fidelity. *Psychosocial Intervention*, 30(*3*), 139 - 153. https://doi.org/10.5093/pi2021a4

I acknowledge that my contribution to the paper is 60%. [redacted]
Asghar Ahmadi

I acknowledge that my contribution to the paper is 8%. [redacted]
Michael Noetel

I acknowledge that my contribution to the paper is 3%. [redacted]
Melissa Schellekens

I acknowledge that my contribution to the paper is 4%. [redacted]
Philip Parker

I acknowledge that my contribution to the paper is 3%. [redacted]
Devan Antczak

I acknowledge that my contribution to the paper is 3%. [redacted]
Mark Beauchamp

I acknowledge that my contribution to the paper is 3%. [redacted]
Theresa Dicke

I acknowledge that my contribution to the paper is 3%. [redacted]
Carmel Diezmann

I acknowledge that my contribution to the paper is 3%. [redacted]
Anthony Maeder

I acknowledge that my contribution to the paper is 3%. [redacted]

Nikos Ntoumanis

I acknowledge that my contribution to the paper is 3%. [redacted]

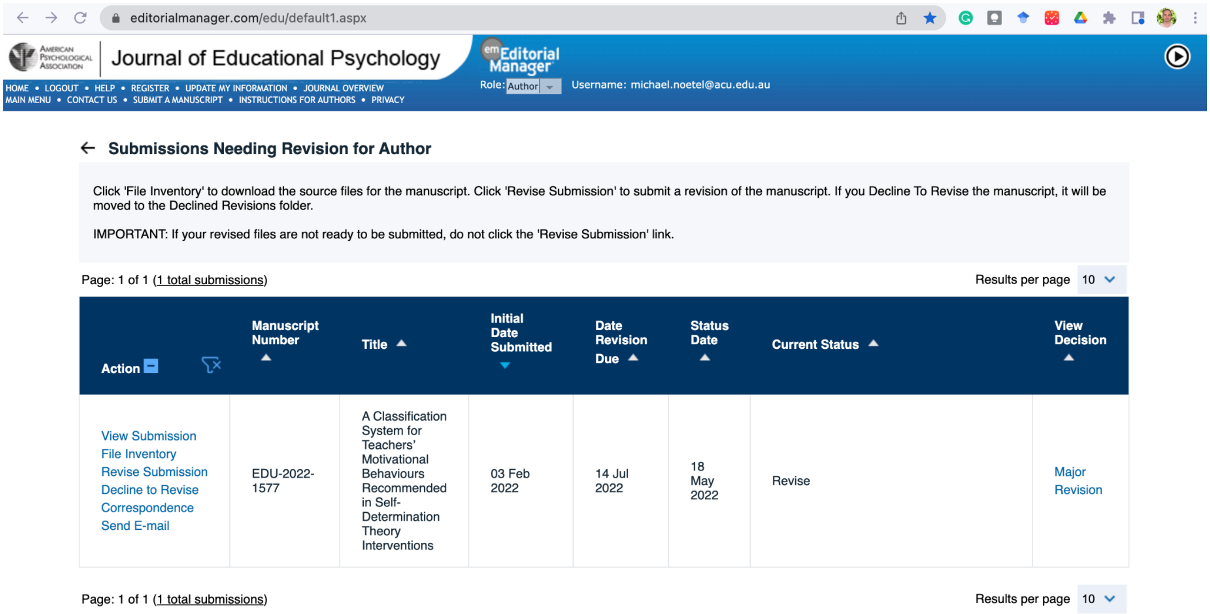
Alexander Yeung

I acknowledge that my contribution to the paper is 4%. [redacted]

Chris Lonsdale

Classification of TMBs Study (Chapter 4)

The classification of teacher motivational behaviours in Chapter 4 has received a revise and resubmit decision in the Journal of Educational Psychology and is currently under review.


editorialmanager.com/edu/default1.aspx

American Psychological Association | Journal of Educational Psychology | Editorial Manager

HOME • LOGOUT • HELP • REGISTER • UPDATE MY INFORMATION • JOURNAL OVERVIEW
MAIN MENU • CONTACT US • SUBMIT A MANUSCRIPT • INSTRUCTIONS FOR AUTHORS • PRIVACY

Role: Author Username: michael.noetel@acu.edu.au

Submissions Needing Revision for Author

Click 'File Inventory' to download the source files for the manuscript. Click 'Revise Submission' to submit a revision of the manuscript. If you Decline To Revise the manuscript, it will be moved to the Declined Revisions folder.

IMPORTANT: If your revised files are not ready to be submitted, do not click the 'Revise Submission' link.

Page: 1 of 1 (1 total submissions) Results per page 10

Action	Manuscript Number	Title	Initial Date Submitted	Date Revision Due	Status Date	Current Status	View Decision
View Submission File Inventory Revise Submission Decline to Revise Correspondence Send E-mail	EDU-2022-1577	A Classification System for Teachers' Motivational Behaviours Recommended in Self-Determination Theory Interventions	03 Feb 2022	14 Jul 2022	18 May 2022	Revise	Major Revision

Page: 1 of 1 (1 total submissions) Results per page 10

I acknowledge that my contribution to the paper is 55%. [redacted]

Asghar Ahmadi

I acknowledge that my contribution to the paper is 5%. [redacted]

Michael Noetel

I acknowledge that my contribution to the paper is 2%. [redacted]

Philip Parker

I acknowledge that my contribution to the paper is 1%. [redacted]

Richard M. Ryan

I acknowledge that my contribution to the paper is 1%. [redacted]

Nikos Ntoumanis

I acknowledge that my contribution to the paper is 1%. [redacted]

Johnmarshall Reeve

I acknowledge that my contribution to the paper is 1%. [redacted]

Mark Beauchamp

I acknowledge that my contribution to the paper is 1%. [redacted]

Theresa Dicke

I acknowledge that my contribution to the paper is 1%. [redacted]

Alexander Yeung

I acknowledge that my contribution to the paper is 1%. [redacted]

Malek Ahmadi

I acknowledge that my contribution to the paper is 1%. [redacted]

Kimberley Bartholomew

I acknowledge that my contribution to the paper is 1%. [redacted]

Thomas K.F. Chiu

I acknowledge that my contribution to the paper is 1%. [redacted]

Thomas Curran

I acknowledge that my contribution to the paper is 1%. [redacted]

Gokce Erturan

I acknowledge that my contribution to the paper is 1%. [redacted]

Barbara Flunger

I acknowledge that my contribution to the paper is 1%. [redacted]

Christina Frederick

I acknowledge that my contribution to the paper is 1%. [redacted]

John Mark Froiland

I acknowledge that my contribution to the paper is 1%.

[redacted]

David González-Cutre

I acknowledge that my contribution to the paper is 1%.

[redacted]

Leen Haerens

I acknowledge that my contribution to the paper is 1%.

[redacted]

Lucas Matias Jeno

I acknowledge that my contribution to the paper is 1%.

[redacted]

Andre Koka

I acknowledge that my contribution to the paper is 1%.

[redacted]

Christa Krijgsman

I acknowledge that my contribution to the paper is 1%.

[redacted]

Jody Langdon

I acknowledge that my contribution to the paper is 1%.

[redacted]

Rhiannon Lee White

I acknowledge that my contribution to the paper is 1%.

[redacted]

David Litalien

I acknowledge that my contribution to the paper is 1%.

[redacted]

David Lubans

I acknowledge that my contribution to the paper is 1%.

[redacted]

John Mahoney

I acknowledge that my contribution to the paper is 1%.

[redacted]

Ma. Jenina N. Nalipay

I acknowledge that my contribution to the paper is 1%.

[redacted]

Erika Patall

I acknowledge that my contribution to the paper is 1%.

[redacted]

Dana Perlman

I acknowledge that my contribution to the paper is 1%.

[redacted]

Eleanor Quested

I acknowledge that my contribution to the paper is 1%.

[redacted]

Sascha Schneider

I acknowledge that my contribution to the paper is 1%.

[redacted]

Martyn Standage

I acknowledge that my contribution to the paper is 1%.

[redacted]

Kim Stroet

I acknowledge that my contribution to the paper is 1%.

[redacted]

Damien Tessier

I acknowledge that my contribution to the paper is 1%.

[redacted]

Cecilie Thogersen-Ntoumani

I acknowledge that my contribution to the paper is 1%.

[redacted]

Henri Tilga

I acknowledge that my contribution to the paper is 1%.

[redacted]

Diego Vasconcellos

I acknowledge that my contribution to the paper is 3%.

[redacted]

Chris Lonsdale

Ethics Approvals

Ethics Approval - Dictionary of Motivational Phrases

For the dictionary study, I used data from a previous longitudinal study entitled "Engaging students during the early years of secondary school: a scalable and sustainable teacher professional learning intervention". The ethics clearance for this project was obtained from the Human Research Ethics Committee of the Australian Catholic University (Approval Number: 2016-198H). Below, please see the clearance email.

Ms Pratigya Pozniak
2016-198H Modification approved
To: Prof. Chris Lonsdale, Dr Ryan Hulteen, Cc: Ms Pratigya Pozniak

New contact info found in this email: Ms Pratigya Pozniak pratigya.pozniak@acu.edu.au

Dear Chris

Ethics Register Number : 2016-198H
Project Title : Engaging Students during the Early Years of Secondary School: A Scalable and Sustainable Teacher Professional Learning Intervention
End Date : 31/12/2021

Thank you for submitting the request to modify form for the above project.

The Chair of the Human Research Ethics Committee has approved the following modification(s):
- Addition of personnel: Mr Asghar Ahmadi

We wish you well in this ongoing research project.

Kind regards,
Ms Pratigya Pozniak

Research Ethics Officer l Office of the Deputy Vice-Chancellor (Research)
Australian Catholic University
T: 02 9739 2646 E: res.ethics@acu.edu.au

THIS IS AN AUTOMATICALLY GENERATED RESEARCHMASTER EMAIL

Ethics Approval - Classification of TMBs Study

The questionnaire and methodology for my third study was approved by the Human Research Ethics committee of the Australian Catholic University (Approval Number: 2020-160E). Below, please see the clearance email.

Asghar Ahmadi

From: Evshen Okan <Evshen.Okan@acu.edu.au> on behalf of Res Ethics <Res.Ethics@acu.edu.au>
Sent: Monday, September 21, 2020 9:19 AM
To: Asghar Ahmadi; Chris Lonsdale
Cc: Res Ethics
Subject: [2020-160E] - Ethics application approved!
Attachments: ACU Research - COVID-19 Update-2.eml

Please review the attached COVID-19 notification as it may contain important information relevant to your research.
Dear Applicant,

Chief Investigator: Professor Chris Lonsdale
Student Researcher: Ahmadi, Asghar
Ethics Register Number: 2020-160E
Project Title: A Classification of Motivation and Behaviour Change Techniques Used in Self-Determination Theory-Based Interventions in Education
Date Approved: 21/09/2020
End Date: 30/09/2021

This is to certify that the above human ethics application has been reviewed by the Australian Catholic University Human Research Ethics Committee (ACU HREC). The application has been approved for the period given above.

Continued approval of this research project is contingent upon the submission of an annual progress report which is due on/before each anniversary of the project approval. A final report is due upon completion of the project. A report proforma can be downloaded from the ACU Research Ethics website.

Researchers are responsible for ensuring that all conditions of approval are adhered to and that any modifications to the protocol, including changes to personnel, are approved prior to implementation. In addition, the ACU HREC must be notified of any reportable matters including, but not limited to, incidents, complaints and unexpected issues.

Researchers are also responsible for ensuring that they adhere to the requirements of the National Statement on Ethical Conduct in Human Research, the Australian Code for the Responsible Conduct of Research and the Universities Research Code of Conduct.

Any queries relating to this application should be directed to the Ethics Secretariat (res.ethics@acu.edu.au). Please quote your ethics approval number in all communications with us.

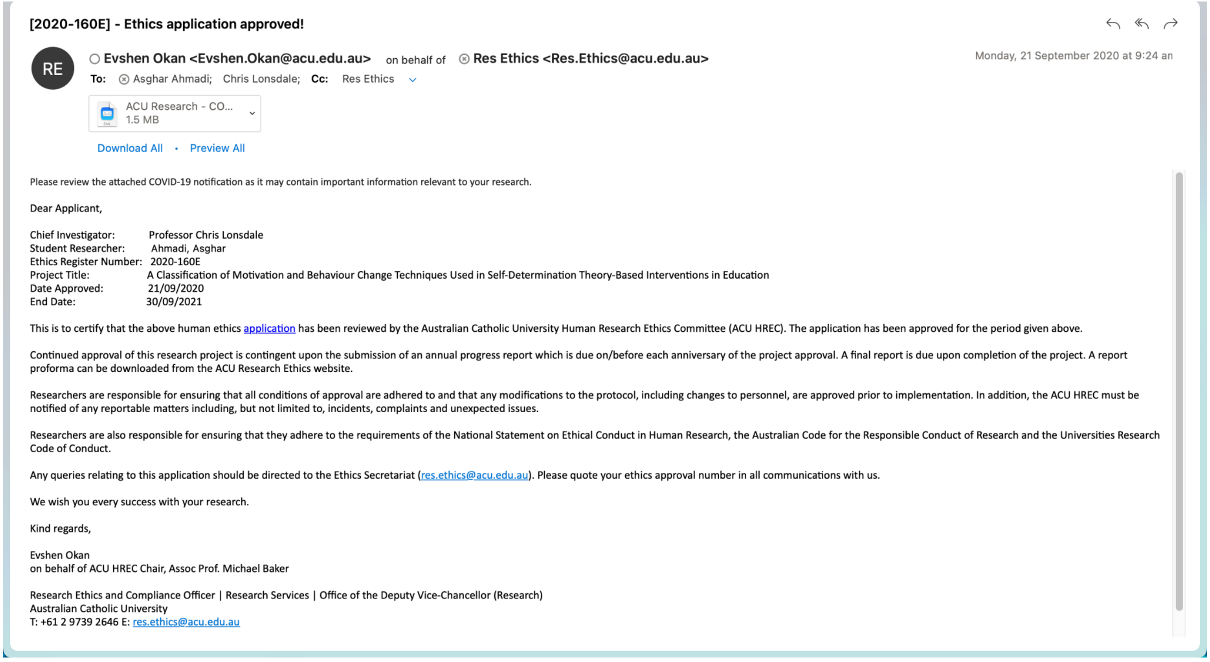
We wish you every success with your research.

Kind regards,

Evshen Okan
on behalf of ACU HREC Chair, Assoc Prof. Michael Baker

Research Ethics and Compliance Officer | Research Services | Office of the Deputy Vice-Chancellor (Research) Australian Catholic University
T: +61 2 9739 2646 E: res.ethics@acu.edu.au

THIS IS AN AUTOMATICALLY GENERATED RESEARCHMASTER EMAIL



[2020-160E] - Ethics application approved!

Evshen Okan <Evshen.Okan@acu.edu.au> on behalf of Res Ethics <Res.Ethics@acu.edu.au>
Monday, 21 September 2020 at 9:24 am
To: Asghar Ahmadi; Chris Lonsdale; Cc: Res Ethics

ACU Research - CO... 1.5 MB

Download All • Preview All

Please review the attached COVID-19 notification as it may contain important information relevant to your research.

Dear Applicant,

Chief Investigator: Professor Chris Lonsdale
Student Researcher: Ahmadi, Asghar
Ethics Register Number: 2020-160E
Project Title: A Classification of Motivation and Behaviour Change Techniques Used in Self-Determination Theory-Based Interventions in Education
Date Approved: 21/09/2020
End Date: 30/09/2021

This is to certify that the above human ethics application has been reviewed by the Australian Catholic University Human Research Ethics Committee (ACU HREC). The application has been approved for the period given above.

Continued approval of this research project is contingent upon the submission of an annual progress report which is due on/before each anniversary of the project approval. A final report is due upon completion of the project. A report proforma can be downloaded from the ACU Research Ethics website.

Researchers are responsible for ensuring that all conditions of approval are adhered to and that any modifications to the protocol, including changes to personnel, are approved prior to implementation. In addition, the ACU HREC must be notified of any reportable matters including, but not limited to, incidents, complaints and unexpected issues.

Researchers are also responsible for ensuring that they adhere to the requirements of the National Statement on Ethical Conduct in Human Research, the Australian Code for the Responsible Conduct of Research and the Universities Research Code of Conduct.

Any queries relating to this application should be directed to the Ethics Secretariat (res.ethics@acu.edu.au). Please quote your ethics approval number in all communications with us.

We wish you every success with your research.

Kind regards,

Evshen Okan
on behalf of ACU HREC Chair, Assoc Prof. Michael Baker

Research Ethics and Compliance Officer | Research Services | Office of the Deputy Vice-Chancellor (Research)
Australian Catholic University
T: +61 2 9739 2646 E: res.ethics@acu.edu.au

Appendices

Appendix A (Chapter 2) - Systematic Review Supplementary Material

Appendix A.1

Automated Coding Models' Description

Model*	Description
Support Vector Machine (SVM)	SVM is a discriminative classifier that is based on the idea of finding a hyperplane that best separates a dataset into two classes. In other words, given the labelled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In a two-dimensional space, this hyperplane is a line dividing a plane into two parts in which each class lies on either side. The learning of the hyperplane in linear SVM is done by transforming the problem using some linear algebra.
Artificial Neural Networks (ANN)	ANN is an algorithm that mimics how the human brain processes information and consist of input and output layers, as well as (in most cases) hidden layers. This model handles the regression and classification problems without the need to explicitly specify any relationships between the input and output variables. ANN model iterates the feedforward and backpropagation processes to identify the optimal amounts of the weights of the network for a single output so the difference between the predicted and the observed outputs is as small as possible.
Markov Chain	Markov chain is a stochastic model describing a sequence of possible events. The process of calculating the probability of each event depends only on the state attained in the previous event and not the sequence of states.
Hidden Markov (HMM)	HMM is an evolved version of the Markov Chain model and assigns labels to each unit in a sequence that are observable or not observable in the world. However, some events such as part-of-speech tags or acoustic events are not observable in the world, so they are called “hidden”. HMM computes a probability distribution over possible labels and chooses the best label sequence.
Maximum Entropy (MaxEnt)	MaxEnt (also known as multinomial logistic regression) is a machine learning framework and belongs to a family of classifiers known as “exponential” or “log-linear” classifiers. This model is used for sequence labelling or classification (i.e., assign labels to each event in some sequences). This model is also capable of assigning a weight to particular events. The most common MaxEnt classifier is Maximum Entropy Markov Model.
Maximum Entropy Markov (MEMM)	MEMM is a graphical model for sequence labelling that combines features of Hidden Markov (HMM) and Maximum Entropy (MaxEnt) models. MEMMs model are applied in Natural language processing, specifically in part-of-speech-tagging and information extraction.
Decision Tree and J48	A decision tree is a hierarchical decision model and consist of “if” and “else” questions asked in each node. Eventually, these questions and the path will lead to a predicted class or a continues real-valued outcome. J48 (C4.5) is an algorithm used to generate a decision tree prediction model. This

	algorithm was developed after ID3 (which is also a decision tree classifier) by Ross Quinlan. The important features of this model are the prediction of discrete and real-valued outcomes, managing missed-values in the input dataset, the prunning ability to prevent overfitting, and weighting features.
Random Forest	Random Forest model is an ensemble learning model for regression, classification and other tasks. This model applies some decision tree models to predict the outcomes and uses the most common prediction or mean of the predictions as the final outcome prediction for each observation.
Conditional Random Field (CRF)	CRFs are a class of statistical modelling method used to predict sequences rather than discrete or real-valued outcomes. This model is best suited to prediction where contextual information or state of the neighbours affect the current prediction.
Labelled Topic	A labelled Topic model is a statistical model that discovers the abstract topics that occur in a series of documents. This model is mainly used in natural language processing and text mining applications.
Latent Dirichlet Allocation (LDA) and DiscLDA	LDA is a generative probabilistic model used for topic modelling purposes. This model is capable of discovering the hidden topics in a corpus, classify documents based on those topics and summarise corpora in terms of the topics identified. DiscLDA is a discriminative variation on Latent Dirichlet Allocation (LDA) model in which a class-dependent linear transformation is introduced on the topic mixture proportions.
Maximum Likelihood	The goal of the Maximum Likelihood model is to find the optimal way to fit a distribution to the data. Based on the data, the distribution can be normal, exponential, gamma or other distributions. This model estimates the parameters of a statistical model given observations, by finding the parameter values that maximize the likelihood of making the observations given the parameters.
AdaBoost	AdaBoost (Adaptive Boosting) is a type of ensemble learning method which uses an iterative approach to learn from the mistakes of weak classifiers to build a stronger learning algorithm. The basic classifiers could be any classifier, from Decision Trees to Logistic Regression.
Automated Co-occurrence Analysis for Semantic Mapping (ACASM)	ACASM constructs a map of the text in terms of thematic nuclei active in it. It works through invariant, ostensible, yet context-sensitive procedures, defined in terms of computational algorithms.
Boostexter	BoosTexter is a text-mining tool that uses a machine-learning technique named boosting (using variations of AdaBoost algorithm). This model categorises a text corpus by combining many simple and moderately inaccurate categorization rules into a single, highly accurate categorization rule.
Discourse Flow Analysis (DFA)	DFA is a technique developed specifically for the psychotherapy domain and focuses on temporal patterns of meanings rather than on the survey of discrete contents. It also considers the contextual features.
Discursis software	Discursis is an automated computer visualisation measurement software that is used to analyse conversational behaviour. Discursis automatically builds an

	internal language model from a transcript, mines the transcript for its conceptual content, and generates an interactive visual account of the discourse. The resultant visual account of the whole consultation can be analysed for patterns of engagement between interactants.
Fidelity Automatic RatEr (FARE)	FARE is a computational system that uses a transcribed text as input, then applies a Decision Tree algorithm that categorizes linguistic patterns associated with high or low fidelity.
K-Nearest Neighbors	K-Nearest-Neighbours is a simple algorithm that classifies events based on a similarity measure (e.g., distance functions). This model classifies the events based on the class that is most common among K neighbours of that event. K is decided by the researcher.
Linear Regression	Given a set of observations, each observation is associated with some features. Linear Regression model is used to predict some real-valued outcome for each observation. The predictive power of this model is boosted when more than one feature is used (in this situation, the model is called multiple linear regression).
Logistic Regression and Lasso Logistic Regression and Ridge Logistic Regression	Given a set of observations, each observation is associated with some features. The Logistic Regression model is used to predict some discrete outcomes such as classes. For example, in a binary classification of cancer detection, considering some features, the model outcomes are two classes: “positive” or “negative”. In case the model is used to predict many discrete outcomes, the model is called “multinomial logistic regression”. Least Absolute Shrinkage and Selection Operator (LASSO) logistic regression model is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. In Ridge Logistic Regression model, variables with minor contribution have their coefficients close to zero. However, all the variables are incorporated in the model. This is useful when all variables need to be incorporated in the model according to domain knowledge.
Naive Bayes	Naive Bayes model is a classification method based on Bayes Theorem. This model assumes that the predictors of a feature are independent from each other. This classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. The three variations of this model are Gaussian, Multinomial, and Bernoulli. Gaussian assumes that the features follow a normal distribution, Multinomial model is used for classification problems with discrete features (e.g., word counts for text classification). In the Bernoulli model, the features are assumed to be a binary-valued (Bernoulli, boolean) variable.
RapidMiner	RapidMiner is a text mining program. This program is suitable for those who are not interested in programming or simply prefer to use the existing software. Developers of this program tried to integrate various operations in the field of data science which allows researchers to quickly apply it for data mining operations.

Note. *We described the main methods in this table. Some of the included papers used variations of these methods. We reported the specified methods used in each paper in Table 1, Automated Coding Method column.

Descriptions of Coding models’ Accuracy Measures

Cohen's kappa. Cohen's kappa coefficient (k) is used for measuring the agreement between human coders and a coding model's prediction. It is computed as (observed accuracy—expected accuracy)/(1—expected accuracy). Cohen's kappa value ranges between 0-1. Kappas <0.40 are considered "fair" to "poor," 0.41–0.60 are "moderate," 0.61–0.80 are "substantial," and >0.81 are "almost perfect" (Landis & Koch, 1977).

Inta-Cater Correlation Coefficient. The Intraclass Correlation Coefficient (ICC) is a measure of the reliability of measurements or ratings. This measure shows the agreement between a human coder and a model on a session-level prediction. An ICC<0.40 is considered a poor level of agreement, ICC between 0.40-0.59 is a fair agreement, ICC between 0.60-0.74 is a good agreement and ICC between 0.75-1.00 is considered an excellent level of agreement (Cicchetti, 1994).

Confusion Matrix. Confusion Matrix describes the complete performance of the model and presents the exact number of the codes that a model predicted correctly or incorrectly.

Confusion Matrix

		Actual Class	
		Positive	Negative
Predicted Class	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Receiver Operating Characteristic (ROC). The Receiver Operating Characteristic (ROC) curve is a plot which shows the performance of a binary classifier as a function of its cut-off threshold. It essentially shows the true positive rate (TPR) against the false positive rate (FPR) for various threshold values.

Area Under the Curve (AUC). AUC calculates the area under the ROC curve, and therefore it is between 0 and 1. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example. AUC can be ranged between 0 to 1. An $AUC=0.5$ indicates a prediction better than chance level, $0.5<AUC<0.7$ is a poor prediction, $0.7=<AUC<0.8$ is an acceptable level of prediction, $0.8=<AUC<0.9$ is an excellent prediction and $AUC>=0.9$ is an outstanding prediction (Hosmer, Lemeshow, & Sturdivant, 2013).

Accuracy. Accuracy is one of the easiest predictive measures and can simply be calculated as the proportion of correctly classified codes (TP+TN) to all the predicted codes (TP+FP+FN+TN). Accuracy is a good measure when the target variable classes in the data are nearly balanced. Accuracy is a good measure when the target variable classes in the data are nearly balanced.

Precision. Precision is a measure that shows what proportion of "positive" predicted codes are actually positive. It is calculated by the number of True Positives divided by the total number of the "Positive" predicted codes (TP+FP).

Recall or Sensitivity. Recall or Sensitivity shows what proportion of actual positives are predicted correctly. It is calculated by the number of True Positives divided by the truly predicted positives and falsely predicted negatives (TP+FN).

Specificity. Specificity is a measure that shows what proportion of actual negatives, were predicted by the model as negative. It is calculated by this formula: TN/FP+TN. Specificity is the exact opposite of Recall.

F1-Score. F1-score is the harmonic mean of precision and recall and is calculated using this formula: F1-score=Precision+Recall/2*Precision*Recall.

Appendix A.2

Search Strategy

Keywords to search in databases

Participants: (teacher* OR coach* OR nurse* OR doctor* OR physician* OR "general practitioner*" OR surgeon* OR psychiatrist* OR interviewer* OR clinician* OR therap* OR dentist* OR physiotherapist* OR chiropractor* OR psychotherap* OR psychologist* OR counselor* OR counsellor* OR "social worker*" OR "care provider*")

Measurement: (analys* OR analyz* OR asses* OR evaluat* OR classif* OR coding OR code OR coded OR rating OR rate* OR annotat*)

Automated coding method: ("language processing" OR "artificial intelligence" OR "text min*" OR "data min*" OR "machine learning" OR "automatic prediction" OR "text classif*" OR "markov model*" OR "topic model*" OR "entropy model*" OR "language model*" OR "computational intelligence" OR "recursive neural network*" OR "discrete sentence feature*" OR "latent dirichlet allocation" OR "long short term memory")

Type of behaviour: (behavio* OR interaction* OR interpersonal OR relation* OR communication* OR conversation* OR fidelity OR integrity)

We conducted the main search on 26th Feb 2019.

We updated the search on 21st Feb 2021.

Database Search Strategy

PubMed Search Strategy

PubMed MeSH terms:

Participants: Educational Personnel OR Health Personnel OR Health Occupations

Measurement: -

Automated coding method: Natural Language Processing OR Artificial Intelligence

Type of the behaviour: Verbal Behavior OR Interpersonal Relations OR Professional-Patient Relations

Main search terms in title and abstract, MeSH terms in "MeSH Major Topic".

Results: 1,586 records

Search update: 1104 records

Education Source Search strategy

Subject headings:

Participants: Teachers OR coaches OR Health occupations OR Interviewers OR Allied health personnel OR Social workers

Measurement: Interaction analysis in education OR Teacher evaluation standards

Automated coding method: Artificial intelligence OR Algorithms

Type of behaviour: Verbal behavior OR Interpersonal relations OR Interpersonal communication OR Teacher-student relationships OR Teacher-student communication

Main search terms in title and abstract, Subject Headings in "Subject".

Results: 198 records

Search update: 81 records

CINAHL Complete Search Strategy

Subject headings:

Participants: Health Personnel OR Social Workers

Measurement: -

Automated coding method: Natural Language Processing OR Artificial Intelligence

Type of behaviour: Interpersonal Relations OR Verbal Behavior

Main search terms in title and abstract, Subject Headings in “Exact Major Subject Heading”.

Results: 297 records

Search update: 271 records

ERIC Search Strategy

Thesaurus:

Participants: Health Occupations OR Health Personnel

Measurement: Counselor Evaluation OR Classification OR Teacher Evaluation

Automated coding method: Natural Language Processing OR Artificial Intelligence

Type of behaviour: Interpersonal Relationship OR Verbal Communication OR Interpersonal Communication

Main search terms in title and abstract, Subject Headings in “Descriptors (exact)”.

Results: 161 records

Search update: 32 records

PsycINFO Search Strategy

Subject headings:

Participants: Educational Personnel OR Health Personnel OR Professional Personnel OR Social Workers

Measurement: Behavior Analysis OR Classification

Automated coding method: Artificial Intelligence OR Automated Information Coding OR Automated Information Processing

Type of behaviour: Verbal Communication OR Teacher Student Interaction OR Interpersonal Interaction OR Interpersonal Relationships OR Interpersonal Communication

Main search terms in title and abstract, Subject Headings in “MeSH Subject Headings”.

Results: 373 records

Search update: 148 Records

SPORTDiscus search strategy

Subject headings:

Participants: TEACHERS OR “COACHES (Athletics)” OR MEDICAL personnel OR PHYSICAL therapists OR SPORTS psychologists

Measurement: CLASSIFICATION

Automated coding method: -

Type of behaviour: COMMUNICATION in sports OR PHYSICIAN-patient relations

Main search terms in title and abstract, Subject Headings in “Subjects (Descriptors)”.

Results: 18 records

Search update: 11 records

Embase Classic+Embase Search strategy

Subject headings:

Participants: educational personnel OR social worker OR health care personnel

Measurement: -

Automated coding method: natural language processing OR artificial intelligence

Behaviour: -

Main search terms in title and abstract, Subject Headings in “Subject heading”.

Results: 1,645 records

Search update: 602 records

Scopus Search Strategy

Scopus database does not have subject headings

Main search terms in title and abstract.

Results: 5076 records

Search update: 2903 records

Computers & applied sciences Search strategy

Computers & applied sciences Search strategy database does not have subject headings

Main search terms in title and abstract.

Results: 193

Search update: 139 records

Total: 9,547

Total records added in the search update: 5291 records

Included Papers

Althoff, T., Clark, K., & Leskovec, J. (2016). Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. Transactions of the Association for Computational Linguistics, 4, 463–476. https://doi.org/10.1162/tacl_a_00111

Atkins, D., Steyvers, M., Imel, Z., & Smyth, P. (2014). Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. Implementation Science: IS, 9(1), 49. https://doi.org/10.1186/1748-5908-9-49

Blanchard, N., Donnelly, P. J., Olney, A. M., Samei, B., Ward, B., Sun, X., Kelly, S., Nystrand, M., & D'Mello, S. K. (2016a). Semi-Automatic Detection of Teacher Questions from Human-Transcripts of Audio in Live Classrooms. Proceedings of the 9th International Conference on Educational Data Mining. https://eric.ed.gov/?id=ED592742

Blanchard, N., Donnelly, P., Olney, A., Samei, B., Ward, B., Sun, X., Kelly, S., Nystrand, M., & D'Mello, S. K. (2016b). Identifying teacher questions using automatic speech recognition in classrooms. Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 191–201. https://www.aclweb.org/anthology/W16-3623.pdf

Can, D., Atkins, D. C., & Narayanan, S. S. (2015). A dialog act tagging approach to behavioral coding: A case study of addiction counseling conversations. Sixteenth Annual Conference of the International Speech Communication Association. https://www.isca-speech.org/archive/interspeech_2015/i15_0339.html

Can, D., Georgiou, P., Atkins, D., Marín, R., Imel, Z., & Narayanan, S. (2016). "It Sounds

Like ...": A Natural Language Processing Approach to Detecting Counselor Reflections in Motivational Interviewing. Journal of Counseling Psychology, 63(3), 343–350. https://doi.org/10.1037/cou0000111

Can, D., Georgiou, P. G., & Atkins, D. C. (2012). A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features. Annual Conference of …. https://www.isca-speech.org/archive/interspeech_2012/i12_2254.html

Cao, J., Tanana, M., Imel, Z., Poitras, E., Atkins, D., & Srikumar, V. (2019). Observing Dialogue in Therapy: Categorizing and Forecasting Behavioral Codes. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 5599–5611. https://doi.org/10.18653/v1/P19-1563

Chakravarthula, S. N., Xiao, B., Imel, Z. E., Atkins, D. C., & Georgiou, P. G. (2015). Assessing Empathy Using Static and Dynamic Behavior Models Based on Therapist's Language in Addiction Counseling. Sixteenth Annual Conference of the International Speech Communication Association. https://doi.org/10.7287/peerj-cs.59v0.1/reviews/3

Chen, Z., Singla, K., Gibson, J., Can, D., Imel, Z. E., Atkins, D. C., Georgiou, P., & Narayanan, S. (2019). Improving the Prediction of Therapist Behaviors in Addiction Counseling by Exploiting Class Confusions. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6605–6609. https://doi.org/10.1109/ICASSP.2019.8682885

Donnelly, P. J., Blanchard, N., Olney, A. M., Kelly, S., Nystrand, M., & D'Mello, S. K. (2017). Words matter: automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. Proceedings of the Seventh International Learning Analytics & Knowledge Conference, 218–227.

https://doi.org/10.1145/3027385.3027417

Donnelly, P. J., Blanchard, N., Samei, B., Olney, A. M., Sun, X., Ward, B., Kelly, S., Nystrand, M., & D’Mello, S. K. (2016). Multi-sensor modeling of teacher instructional segments in live classrooms. Proceedings of the 18th ACM International Conference on Multimodal Interaction, 177–184. https://doi.org/10.1145/2993148.2993158

Donnelly, P. J., Blanchard, N., Samei, B., Olney, A. M., Sun, X., Ward, B., Kelly, S., Nystran, M., & D’Mello, S. K. (2016). Automatic Teacher Modeling from Live Classroom Audio. Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization, 45–53. https://doi.org/10.1145/2930238.2930250

Flemotomos, N., Martinez, V. R., Gibson, J., Atkins, D. C., Creed, T., & Narayanan, S. S. (2018). Language Features for Automated Evaluation of Cognitive Behavior Psychotherapy Sessions. INTERSPEECH, 1908–1912. 10.21437/Interspeech.2018-1518

Gallo, C., Pantin, H., Villamar, J., Prado, G., Tapia, M., Ogihara, M., Cruden, G., & Brown, C. H. (2015). Blending Qualitative and Computational Linguistics Methods for Fidelity Assessment: Experience with the Familias Unidas Preventive Intervention. Administration and Policy in Mental Health, 42(5), 574–585. https://doi.org/10.1007/s10488-014-0538-4

Gaut, G., Steyvers, M., Imel, Z., Atkins, D., & Smyth, P. (2017). Content Coding of Psychotherapy Transcripts Using Labeled Topic Models. IEEE Journal of Biomedical and Health Informatics, 21(2), 476–487. https://doi.org/10.1109/JBHI.2015.2503985

Gibson, J., Atkins, D., Creed, T., Imel, Z., Georgiou, P., & Narayanan, S. (2019). Multi-label Multi-task Deep Learning for Behavioral Coding. IEEE Transactions on Affective Computing, 1–1. https://doi.org/10.1109/TAFFC.2019.2952113

Gibson, J., Can, D., Georgiou, P., Atkins, D. C., & Narayanan, S. S. (2017). Attention Networks for Modeling Behaviors in Addiction Counseling. Interspeech 2017, 3251–3255. https://doi.org/10.21437/Interspeech.2017-218

Gibson, J., Can, D., Xiao, B., Imel, Z. E., Atkins, D. C., Georgiou, P., & Narayanan, S. S. (2016). A Deep Learning Approach to Modeling Empathy in Addiction Counseling. Interspeech 2016, 2016, 1447–1451. https://doi.org/10.21437/Interspeech.2016-554

Goldberg, S. B., Flemotomos, N., & Martinez, V. R. (2020). Machine learning and natural language processing in psychotherapy research: Alliance as example use case. Journal of Counseling. https://psycnet.apa.org/record/2020-46941-003

Gupta, R., Georgiou, P. G., Atkins, D. C., & Narayanan, S. S. (2014). Predicting client's inclination towards target behavior change in motivational interviewing and investigating the role of laughter. Fifteenth Annual Conference of the International Speech Communication Association. https://www.isca-speech.org/archive/interspeech_2014/i14_0208.html

Hasan, M., Carcone, A. I., Naar, S., Eggly, S., Alexander, G. L., Hartlieb, K. E. B., & Kotov, A. (2019). Identifying Effective Motivational Interviewing Communication Sequences Using Automated Pattern Analysis. Journal of Healthcare Informatics Research, 3(1), 86–106. https://doi.org/10.1007/s41666-018-0037-6

Hasan, M., Kotov, A., Carcone, A., Dong, M., Naar, S., & Hartlieb, K. B. (2016). A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories. Journal of Biomedical Informatics, 62, 21–31. https://doi.org/10.1016/j.jbi.2016.05.004

Hasan, M., Kotov, A., Carcone, A. I., Dong, M., & Naar, S. (2018). Predicting the Outcome of Patient-Provider Communication Sequences using Recurrent Neural Networks and Probabilistic Models. AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science, 2017, 64–73. https://www.ncbi.nlm.nih.gov/pubmed/29888043

Howes, C., Purver, M., & McCabe, R. (2013). Using conversation topics for predicting therapy outcomes in schizophrenia. Biomedical Informatics Insights, 6(Suppl 1), 39–50. https://doi.org/10.4137/BII.S11661

Idalski Carcone, A., Hasan, M., Alexander, G. L., Dong, M., Eggly, S., Brogan Hartlieb, K., Naar, S., MacDonell, K., & Kotov, A. (2019). Developing Machine Learning Models for Behavioral Coding. Journal of Pediatric Psychology, 44(3), 289–299. https://doi.org/10.1093/jpepsy/jsy113

Imel, Z., Steyvers, M., & Atkins, D. (2015). Computational psychotherapy research: Scaling up the evaluation of patient-provider interactions. Psychotherapy , 52(1), 19–30. https://doi.org/10.1037/a0036841

Lacson, R., & Barzilay, R. (2005). Automatic processing of spoken dialogue in the home hemodialysis domain. AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium, 420–424. https://www.scopus.com/inward/record.uri?eid=2-s2.0-39049191868&partnerID=40&md5=c99769d457f6ce7d55d48949fa865b04

Malandrakis, N., & Narayanan, S. S. (2015). Therapy language analysis using automatically generated psycholinguistic norms. Sixteenth Annual Conference of the International Speech Communication Association. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84959123480&partnerID=40&md5=48c4f18021e9a3d03bf0493b5f6ee649

Mayfield, E., Laws, M. B., Wilson, I. B., & Penstein Rosé, C. (2014). Automating annotation of information-giving for analysis of clinical conversation. Journal of the American Medical Informatics Association: JAMIA, 21(e1), e122–e128. https://doi.org/10.1136/amiajnl-2013-001898

Mieskes, M., & Stiegelmayr, A. (2019). Preparing data from psychotherapy for natural language processing (Isahara H., Maegaard B., Piperidis S., Cieri C., Declerck T., Hasida K., Mazo H., Choukri K., Goggi S., Mariani J., Moreno A., Calzolari N., Odijk J., & Tokunaga T. (eds.); pp. 2896–2902). European Language Resources Association (ELRA). https://www.scopus.com/inward/record.uri?eid=2-s2.0-85059902190&partnerID=40&md5=f539dde0dce50ce71931c25747add595

Nitti, M., Ciavolino, E., Salvatore, S., & Gennaro, A. (2010). Analyzing psychotherapy process as intersubjective sensemaking: an approach based on discourse analysis and neural networks. Psychotherapy Research: Journal of the Society for Psychotherapy Research, 20(5), 546–563. https://doi.org/10.1080/10503301003641886

Park, J., Jindal, A., Kuo, P., Tanana, M., Lafata, J. E., Tai-Seale, M., Atkins, D. C., Imel, Z. E., & Smyth, P. (2021). Automated rating of patient and physician emotion in primary care visits. Patient Education and Counseling. https://doi.org/10.1016/j.pec.2021.01.004

Park, J., Kotzias, D., Kuo, P., Logan, R. L., Iv, Merced, K., Singh, S., Tanana, M., Karra Taniskidou, E., Lafata, J. E., Atkins, D. C., Tai-Seale, M., Imel, Z. E., & Smyth, P. (2019). Detecting conversation topics in primary care office visits from transcripts of patient-provider interactions. Journal of the American Medical Informatics Association: JAMIA, 26(12), 1493–1504. https://doi.org/10.1093/jamia/ocz140

Pérez-Rosas, V., Mihalcea, R., Resnicow, K., Singh, S., An, L., Goggin, K. J., & Catley, D. (2017). Predicting Counselor Behaviors in Motivational Interviewing Encounters.

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, 1128–1137. https://www.aclweb.org/anthology/E17-1106

Perez-Rosas, V., Sun, X., Li, C., Wang, Y., Resnicow, K., & Mihalcea, R. (n.d.). Analyzing the Quality of Counseling Conversations: the Tell-Tale Signs of High-quality Counseling.

Salvatore, S., Gennaro, A., Auletta, A. F., Tonti, M., & Nitti, M. (2012). Automated method of content analysis: a device for psychotherapy process research. Psychotherapy Research: Journal of the Society for Psychotherapy Research, 22(3), 256–273. https://doi.org/10.1080/10503307.2011.647930

Samei, B., Olney, A. M., Kelly, S., Nystrand, M., & D'Mello, S. (2014). Domain Independent Assessment of Dialogic Properties of Classroom Discourse. Grantee Submission. https://eric.ed.gov/?id=ED566380

Samei, B., Olney, A. M., Kelly, S., Nystrand, M., D'Mello, S., Blanchard, N., & Graesser, A. (2015). Modeling Classroom Discourse: Do Models That Predict Dialogic Instruction Properties Generalize across Populations? International Educational Data Mining Society. https://eric.ed.gov/?id=ED560879

Sen, T., Ali, M. R., Hoque, M. E., Epstein, R., & Duberstein, P. (2017). Modeling doctor-patient communication with affective text analysis. In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). https://doi.org/10.1109/acii.2017.8273596

Singla, K., Chen, Z., Flemotomos, N., Gibson, J., Can, D., Atkins, D., & Narayanan, S. (2018). Using Prosodic and Lexical Information for Learning Utterance-level Behaviors in Psychotherapy. Interspeech 2018, 3413–3417. https://doi.org/10.21437/Interspeech.2018-2551

Song, Y., Lei, S., Hao, T., Lan, Z., & Ding, Y. (2020). Automatic Classification of Semantic Content of Classroom Dialogue. Journal of Educational Computing Research, 0735633120968554. https://doi.org/10.1177/0735633120968554

Suresh, A., Sumner, T., Jacobs, J., Foland, B., & Ward, W. (2019). Automating analysis and feedback to improve mathematics teachers' classroom discourse. Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence, 33, 9721–9728. https://doi.org/10.1609/aaai.v33i01.33019721

Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., & Srikumar, V. (2016). A Comparison of Natural Language Processing Methods for Automated Coding of Motivational Interviewing. Journal of Substance Abuse Treatment, 65, 43–50. https://doi.org/10.1016/j.jsat.2016.01.006

Velasquez, P. A. E., & Montiel, C. J. (2018). Reapproaching Rogers: a discursive examination of client-centered therapy. Person-Centered and Experiential Psychotherapies, 17(3), 253–269. https://doi.org/10.1080/14779757.2018.1527243

Wallace, B. C., Laws, M. B., Small, K., Wilson, I. B., & Trikalinos, T. A. (2014). Automatically annotating topics in transcripts of patient-provider interactions via machine learning. Medical Decision Making: An International Journal of the Society for Medical Decision Making, 34(4), 503–512. https://doi.org/10.1177/0272989X13514777

Wallace, B. C., Trikalinos, T. A., Laws, M. B., Wilson, I. B., & Charniak, E. (2013). A Generative Joint, Additive, Sequential Model of Topics and Speech Acts in Patient-Doctor Communication. Proceedings of the 2013 Conference on Empirical Methods in

Natural Language Processing, 1765–1775. https://www.aclweb.org/anthology/D13-1182

Wang, Z., Pan, X., Miller, K. F., & Cortina, K. S. (2014). Automatic classification of activities in classroom discourse. Computers & Education, 78, 115–123. https://doi.org/10.1016/j.compedu.2014.05.010

Xiao, B., Can, D., Georgiou, P. G., Atkins, D., & Narayanan, S. S. (2012). Analyzing the Language of Therapist Empathy in Motivational Interview based Psychotherapy. Signal and Information Processing Association Annual Summit and Conference (APSIPA), ... Asia-Pacific. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, 2012. https://www.ncbi.nlm.nih.gov/pubmed/27602411

Xiao, B., Can, D., Gibson, J., Imel, Z., Atkins, D., Georgiou, P., & Narayanan, S. (2016). Behavioral Coding of Therapist Language in Addiction Counseling Using Recurrent Neural Networks. Interspeech 2016, 2016, 908–912. https://doi.org/10.21437/Interspeech.2016-1560

Xiao, B., Huang, C., Imel, Z., Atkins, D., Georgiou, P., & Narayanan, S. (2016). A technology prototype system for rating therapist empathy from audio recordings in addiction counseling. PeerJ. Computer Science, 2(4). https://doi.org/10.7717/peerj-cs.59

Xiao, B., Imel, Z. E., Georgiou, P., Atkins, D., & Narayanan, S. (2015). " Rate my therapist": automated detection of empathy in drug and alcohol counseling via speech and language processing. PloS One, 10(12), e0143055. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0143055

Appendix A.3

Predictive Performance of each Method

Model	Study	Predictive accuracy measure	Value	Codes being predicted	Interpretation	Size of dataset
Latent Dirichlet Allocation	Atkins et al., 2014	Area Under the Curve	0.62-0.81	10	Better than Chance = All the codes Poor = 3 codes Acceptable =: 5 codes Excellent = 2 codes Outstanding = 0 codes	1,004,924 words
Latent Dirichlet Allocation	Atkins et al., 2014	Intra-Class Correlation	Excellent = 7 codes Good = 1 code Fair = 0 codes Poor = 2 codes	10	Excellent =7 codes Good =1 code Fair = 0 codes Poor = 2 codes	1,004,924 words
Latent Dirichlet Allocation	Atkins et al., 2014	Kappa	Almost perfect agreement = 1 code Substantial = 5 codes moderate or less = 4 codes	10	Almost perfect agreement = 1 code Substantial = 5 codes moderate or less = 4 codes	1,004,924 words
Maximum Entropy Markov Model	Can et al., 2012	F1-score	0.81	10	--	sessions 57

Hidden Markov Model	Can et al., 2012	F1-score	0.71	10	--	sessions 57
Conditional Random Field	Can et al., 2015	F1-score	MISC28-code = 0.75	19	--	1,736,000 words
Maximum Entropy Markov Model	Can et al., 2016	F1-score	0.81	3	--	sessions 57
Naive Bayes	Carcone et al., 2019	Kappa	0.497	41	Moderate	utterances 11,353
Naive Bayes	Carcone et al., 2019	F1-score	0.55	41	--	utterances 11,353
Naive Bayes-Multinomial	Carcone et al., 2019	Kappa	0.62	41	Substantial	utterances 11,353
Naive Bayes-Multinomial	Carcone et al., 2019	F1-score	0.64	41	--	utterances 11,353
J48	Carcone et al., 2019	Kappa	0.54	41	Moderate	utterances 11,353
J48	Carcone et al., 2019	F1-score	0.58	41	--	utterances 11,353
AdaBoost	Carcone et al., 2019	Kappa	0.57	41	Moderate	utterances 11,353
AdaBoost	Carcone et al., 2019	F1-score	0.61	41	--	utterances 11,353
Random Forest Model	Carcone et al., 2019	Kappa	0.62	41	Substantial	utterances 11,353
Random Forest Model	Carcone et al., 2019	F1-score	0.62	41	--	utterances 11,353
DiscLDA	Carcone et al., 2019	Kappa	0.39	41	Fair	utterances 11,353
DiscLDA	Carcone et al., 2019	F1-score	0.43	41	--	utterances 11,353
Conditional Random Field	Carcone et al., 2019	Kappa	0.51	41	Moderate	utterances 11,353

Conditional Random Field	Carcone et al., 2019	F1-score	0.67	41	--	utterances 11,353
Support Vector Machine	Carcone et al., 2019	Kappa	0.663	41	Substantial	utterances 11,353
Support Vector Machine	Carcone et al., 2019	F1-score	0.68	41	--	utterances 11,353
Static Behavior Model	Chakravarthula et al., 2015	Accuracy	0.81	2	--	sessions 200
Activation-based Dynamic Behavior Model	Chakravarthula et al., 2015	Accuracy	0.755	2	--	sessions 200
Likelihood-based Dynamic Behavior Model	Chakravarthula et al., 2015	Accuracy	0.755	2	--	sessions 200
Fidelity Automatic Rater	Gallo et al., 2015	Pearson Correlation Coefficient	0.32-0.35	3	Weak	86,000 words
Labelled Latent Dirichlet Allocation	Gaut et al., 2017	Area Under the Curve	0.789	41	Acceptable	8,000,000 words
Lasso Logistic Regression	Gaut et al., 2017	Area Under the Curve	0.7	41	Acceptable	8,000,000 words
Deep Neural Networks	Gibson et al., 2016	F1-score	MISC-8 code = 0.643 MISC-28 code = 0.258	19	--	sessions 348
Recurrent Neural Networks with attention-based LSTM	Gibson et al., 2017	F1-score	0.637	8	--	1,659,000 words
Feed-Forward Neural Network	Gibson et al., 2017	F1-score	0.58	8	--	1,659,000 words
Maximum Entropy Markov Model	Gupta et al., 2014	Accuracy	0.7	5	--	sessions 49
Recurrent Neural	Hasan et al., 2018	F1-score	0.86	12	--	sessions

Network						129
Markov Chain Model	Hasan et al., 2018	F1-score	0.7	12	--	sessions 129
Hidden Markov Model	Hasan et al., 2018	F1-score	0.61	12	--	sessions 129
Support Vector Machine	Howes et al., 2013	Accuracy	0.662	20	--	sessions 138
J48	Howes et al., 2013	Accuracy	0.51	20	--	sessions 138
Latent Dirichlet Allocation	Howes et al., 2013	Kolmogorov-Smirnov	D = 0.300, p = 0.257	20	--	sessions 138
Random forest Model	Imel et al., 2015	Accuracy	0.87	4	--	9,300,000 words
Boostexter	Lacson and Barzilay, 2005	Accuracy	0.73	4	--	17,384 words
Linear Regression	Malandrakis and Narayanan, 2015	Pearson Correlation Coefficient	0.8	11	Strong	sessions 312
Ridge Logistic Regression	Mayfield et al., 2014	Accuracy	0.712	3	--	sessions 415
Ridge Logistic Regression	Mayfield et al., 2014	Kappa	0.573	3	Moderate	sessions 415
Random forest Model	Mieskes and Stiegelmayr, 2018	F1-score	0.20-0.46	7	--	sessions 35
Random forest Model	Mieskes and Stiegelmayr, 2018	Kappa	0-0.49	7	Poor to Moderate	sessions 35
Support Vector Machine	Perez-Rosas et al., 2017	F1-score	0.63-0.84	10	--	sessions 277
Support Vector Machine	Perez-Rosas et al., 2019	F1-score	0.87	2	--	sessions 151

Automated Co-occurrence Analysis for Semantic Mapping	Salvatore et al., 2012	Kappa	0.378	14	Fair	sessions 48
K-Nearest-Neighbours	Sen et al., 2017	Accuracy	0.71	N/A	--	sessions 122
Recurrent Neural Network	Singla et al., 2018	F1-score	0.42-0.60	3	--	utterances 85,015
Recurrent Neural Networks	Tanana et al., 2016	F1-score	0-0.93	11	--	1,700,000 words
Discrete Sentence Features	Tanana et al., 2016	F1-score	0-0.94	11	--	1,700,000 words
Discrete Sentence Features	Tanana et al., 2016	Intra-Class Correlation	Excellent = 6 codes Good = 1 codes Fair = 0 codes Poor = 4 codes	11	Excellent = 6 codes Good = 1 codes Fair = 0 codes Poo = 4 codes	1,700,000 words
Recurrent Neural Networks	Tanana et al., 2016	Intra-Class Correlation	Excellent = 6 codes Good: 0 codes Fair = 1 codes Poor = 4 codes	11	Excellent = 6 codes Good = 0 codes Fair = 1 codes Poor = 4 codes	1,700,000 words
Discrete Sentence Features	Tanana et al., 2016	Kappa	Almost Perfect =: 1 codes Substantial = 4 codes Moderate = 1 code	11	Almost Perfect = 1 codes Substantial =4 codes Moderate = 1 code	1,700,000 words

			Fair = 2 codes Slight = 3 codes		Fair = 2 codes Slight = 3 codes	
Recurrent Neural Networks	Tanana et al., 2016	Kappa	Almost Perfect = 1 code Substantial = 1 codes Moderate = 3 codes Fair = 3 codes Slight = 3 codes	11	Almost Perfect = 1 code Substantial = codes Moderate = 3 codes Fair = 3 codes Slight = 3 codes	1,700,000 words
Markov-Multinomial Model	Wallace et al., 2013	F1-score	0.23	33	--	sessions 476
Joint Additive Sequential Model	Wallace et al., 2013	F1-score	0.21	33	--	sessions 476
Conditional Random Field	Wallace et al., 2014	Accuracy	0.64	6	--	sessions 476
Conditional Random Field	Wallace et al., 2014	Kappa	0.47-0.53	6	Moderate	sessions 476
Maximum Likelihood Model	Xiao et al., 2012	F1-score	0.56	2	--	sessions 116
Support Vector Machine	Xiao et al., 2015	F1-score	0.89	2	--	8,298,507 words
Maximum Entropy Markov Model	Xiao, Can, et al., 2016	Accuracy	0.85	2	--	sessions 1,553
Maximum Likelihood Model	Xiao, Can, et al., 2016	Accuracy	0.85	2	--	sessions 1,553

Recurrent Neural Network	Xiao, Huang, et al., 2016	Accuracy	0.75	8	--	1,414,000 words
Maximum Entropy Markov Model	Xiao, Huang, et al., 2016	Accuracy	0.72	8	--	1,414,000 words
Recurrent Neural Networks	Xiao, Huang, et al., 2016b	Kappa	0.40-0.95	8	Moderate	1,414,000 words
Naïve Bayes	Blanchard et al., 2016b (Semi-automatic...)	F1-score	0.66	2		1,000 utterances
J48	Blanchard et al., 2016b (Identifying Teacher...)	F1-score	0.59	2	-	10,080 utterances
J48	Blanchard et al., 2016b (Identifying Teacher...)	Pearson Correlation r	0.85	2	Strong correlation	10,081 utterances
Naïve Bayes classifier using the WEKA machine learning toolbox	Donelly et al., 2016 (Multi-sensor...)	F1-score	0.43-0.51	5	-	32,134 utterances
Naïve Bayes classifier using the WEKA machine learning toolbox	Donnely et al., 2016 (Automatic teacher...)	F1-score	0.23-0.68	5	-	2,254 utterances
J48 Decision Tree	Donnelly et al., 2017	F1-score	0.66	2	-	10,080 utterances
J48 Decision Tree	Samei et al., 2014	Kappa	0.24-0.28	2	Fair	9,579 utterances
J48 Decision Tree	Samei et al., 2014	Accuracy	0.62-0.64	2	-	9,580 utterances
Decision Tree	Samei et al., 2015	Accuracy	0.62-0.68	2	-	20,737 utterances
Random Forest	Wang et al., 2014	Accuracy	0.84	3	-	13 classroom sessions

Random Forest	Wang et al., 2014	Kappa	0.76	3	Substantial	14 classroom sessions
Bidirectional Long Short Term Memory (Bi-LSTM)	Chen et al., 2019	F1-score	0.5876	8	-	sessions 337
Bidirectional Long Short Term Memory (Bi-LSTM)	Chen et al., 2019	Accuracy	0.6328	8	-	sessions 337
Bidirectional Long Short Term Memory (Bi-LSTM)	Gibson et al., 2019	F1-score	0.619	11	-	10,844,000 words
Bidirectional Long Short Term Memory (Bi-LSTM)	Gibson et al., 2019	F1-score	0.784	11	-	10,844,000 words
Logistic Regression	Park et al., 2019	F1-score	0.7515	27	-	talk-turns 122,083
Support Vector Machine	Park et al., 2019	F1-score	0.745	27	-	talk-turns 122,083
Gated Recurrent Unit	Park et al., 2019	F1-score	0.7368	27	-	talk-turns 122,083
Conditional Random Field	Park et al., 2019	F1-score	0.6764	27	-	talk-turns 122,083
HMM-LR	Park et al., 2019	F1-score	0.7655	27	-	talk-turns 122,083
HMM-SVM	Park et al., 2019	F1-score	0.6863	27	-	talk-turns 122,083
HMM-GRU	Park et al., 2019	F1-score	0.7706	27	-	talk-turns 122,083
Hier-GRU	Park et al., 2019	F1-score	0.7778	27	-	talk-turns 122,083
Support Vector Machine	Flemotomos et al., 2018	F1-score	0.86	11	-	sessions 386
Gated Recurrent	Cao et al., 2019	F1-score	0.654	8	-	sessions

Unit						377
Recurrent Neural Network	Park et al., 2021	Pearson correlation coefficient	0.6	3	Moderate	210,000 utterances
Logistic Regression	Park et al., 2021	Pearson correlation coefficient	0.55	3	Moderate	210,000 utterances
CNN-BiLSTM	Song et al., 2020	F1-score	0.68	7	-	sessions 155
Bi-directional Long Short-Term Memory	Suresh et al., 2019	F1-score	0.65	8	-	60,241 sentences
Ridge Regression	Goldberg et al., 2020	Mean Squared Error, Spearman's rank correlation	MSE = 0.67 Spearman's p = 0.15, $p < .001$	2	-	sessions 1235

Appendix B (Chapter 3) - Dictionary Study Supplementary Material

Appendix B.1

Reviewed Built-in Dictionaries in LIWC Program (Pennebaker et al., 2015)

Under the ‘Linguistic Dimensions’ section: ‘Auxiliary verbs’ (e.g., am, will, have), ‘Negations’ (e.g., no, not, never), under the ‘Other Grammar’ section: ‘Comparisons’ (e.g., greater, best, after), under the ‘Psychological Processes’ section: ‘Affective processes’ (e.g., happy, cried), ‘Positive emotion’ (e.g., love, nice, sweet), ‘Negative emotion’ (e.g., hurt, ugly, nasty), ‘Anxiety’ (e.g., worried, fearful), ‘Friends’ (e.g., buddy, neighbour), ‘Insight’ (e.g., think, know), ‘Discrepancy’ (e.g., should, would), ‘Certainty’ (e.g., always, never), ‘Affiliation’ (e.g., ally, friend, social), ‘Achievement’ (e.g., win, success, better), ‘Power’ (e.g., superior, bully), ‘Risk’ (danger, doubt), ‘Swear words’ (e.g., damn, shit), and ‘Assent’ (e.g., agree, OK, yes) dictionaries).

Appendix B.2

Words Collected from the LIWC Dictionaries

Can, could, did, done, may, might, shall, do, don't, must, ought, should, Shouldn't, cannot, couldn't, didn't, cannot, never, nobod*, none, nope, nowhere, should'nt, wasn't, won't, better, brighter, easier, top, funniest, happier, like, faster, luckier, slowest, smallest, worse, worst,advantag*, benefic*, benefit, benefits, benefitt*, desir*, determina*,determined, encourag*, engag*, entertain*, enthus*, excel, favorite, favour*, flexib*, fortunately, free, honest, hope, hopefully, importance, inspir*, intellect*, interest, joy*, keen*, luck, open-minded*, openness, optimal*, passion*, peace, play, popular, positive, profit*, promise*, relax*, reliev*, resolv*, safe, secur*, succes, support, treasur*, triumph*, useful, value, wisdom, worth, admir*, ador*, alright*, amaze*, amazing, amazingly, award*, awesome, beautiful, beautify, beauty, best, better, brave, brilliance*, brilliant, brilliantly, calm, challeng*, champ*, cheer, cheerful, cheers, clever, confidence, confident, creative, decent, dignity, eager, ease*, easier, easily, energ*, excellent, fabulous, fair, fantasi*, fine, glad, glamour, glory, goodness, grace, great, handsome, hero, honor*, impress*, legit, merit*, neat, nice, perfect, pleasant*, polite, prais*, precious*, pride, privileg*, proud, rich, smart, strength*, strong, super, superb*, suprem*, talent*, thrill*, wow*, wise, wonderful, accept*, affection*, agree*, beloved, bonus*, care, cared, caring, certain*, charm*, concerned, considerate, cool, cute, daring, dear, definitely, delicious*, emotion, enjoy*, excited, festiv*, fiesta*, fond, forgiv*, friend, friendly, fun, gorgeous, gratef*, heal, help, hooray, hug, humour*, hurra*, improve*, innocen*, joke*, kindly, laugh*, love, ok, okay, party*, pretty, ready, respect, sorry, sociability, surprise, thank, toleran*, warm, weird, well, yay*, worry, yummy, yum, aggress*, angry, avoid*, credit*, deceiv*, defect*, despair*, destroy*, destruct, destructive, devastat*, disadvantag*, disagree*, disappoint*, discourag*, disgrac*, distress*, disturb*, domina*, dread*, envy*, fault*, fear, fearful*, forbid, forbidden, furious, fury, gloomy, gossip*, grief, guilt, harm, hazard, hazy, hurt*, ideal*, ignore, incentive*, insecur*, insult*, interrup*, intimidat*, irrational*, lose, loss*, lying, mourn*, nervous, nightmar*, obsess*, offence*, offend*, overwhelm*, pain, painf*, panic*, paranoi*, pessimis*, pitiful, pressur*, punish*, regret*, reject*, resign*, rude, ruin*, sad, sarcas*, savage*, serious, shame*, sick, skeptic*, snob, stress*, struggl*, stubborn*, suffer, suspicio*, threat*, tough, trauma*, trick, trivial, troubl*, unaccept*, undesir*, victim*, violat*, vital*, win, worst, yell, annoying, awful, awkward, bad, badly, cheat*, confuse, damn*, danger, dangerous, desperat*, difficult, doubt*, empty, fail*, fake, fatigu*, gross, hilarious, horribly, horrid*, horror*, impatien*, impolite*, incompeten*, indecisive, ineffect*, inferior, jealous, lazy, mess, miser*, poor, powerless*, prize*, reluctan*, selfish*, tedious, uncertain*, tragic, unfair, unfriendly, weak, wrong, alone, worried, afraid, avoid*, confuse, disturb*, fear, guilt, horrid*, horror*, suspicio*, threat*, doubt*, embarrass*, horrible, awkward, shy, tension*, ally, amigo*, bestfriend*, bud, buddies, classmate, crew, darlin*, dude*, fellow, friend, guy*, mate, mates, neighbor*, pal, partner*, peeps, playmate*, schoolmate, sweetheart*, indeed, obviously, exact*, sure*, absolute, absolutely, all, certain*, clear, commitment*, complete, definite, entire*, essential, everything*, everytime, extremely, forever, invariab*, must, necessari*, never, prove*, prefer*, wish, would, yearn*, abnormal*, expect*, if, mistak*, lack, regret*, impossible, inadequa*, accompan*, ally, belong*, brother*, buddies, celebrat*, club, consort*, cooperat*, dance, empath*, fellowship*, joins, let's, member*, union*, advantag*, promot*, abilit*, able, achievable, acquir*, ambition, capab*, champ*, endeav, elit*, leader*, overcome, powerful, practice, progress, team*, burnout*, compet*, dropout*, finaliz*, rank, surviv*, beat, inadequa*, unproduc*, master, best, famous, expert*, lead, lord*, above, allow*, apolog*, control*, forbids, manage, obey, obedient, oppose*, unaccept*, amateur*, bullies, child, kid, abstain*, alarm*, averse, careful*, caution*, cease*, stop, consequen*, danger, prevent*, trust, yield*, bad, fault*, disaster*, bastard*bloody, bullshit, damn*, nigger*, heck, hell, idiot*, piss*, awesome, alright*, cool, whoo*, yaas*, yaaa*, yea, yup

Appendix B.3

Autonomous Versus Controlling Self-talk Dictionary (Oliver et al., 2008)

%	
1	Autonomous
2	Controlling
%	
choice	1
choose	1
chosen	1
chose	1
alternative	1
option	1
could	1
might	1
which	1
may	1
decision	1
opportunity	1
prefer*	1
select*	1
range	1
free*	1
liberty	1
independ*	1
autonom*	1
made	2
ought	2

obliged	2
have	<to>2
got	<to>2
must*	2
should	2
restrict*	2
control*	2
can’t	2
don’t	2
forced	2
unnatural	2
inhibit*	2
repress*	2
held	2

Oliver, E. J., Markland, D., Hardy, J., & Petherick, C. M. (2008). The effects of autonomy-supportive versus controlling environments on self-talk. Motivation and Emotion, 32(3), 200–212. https://doi.org/10.1007/s11031-008-9097-x

Appendix B.4

Frequency and Weighted Log Odds Scores for Dictionary Words

Dictionary	Sub-category	Dictionary words (based on the ratings of judging panel)	Frequency (in training dataset, 44 teachers high/low need supportive)	How many teachers used it? (from 66 teachers in training dataset)	Weighted_log_odds
Need Supportive	Autonomy Supportive				
		might	148	52	-0.352
		because	469	65	-0.543
		could	152	46	-0.522
		choice	12	15	-0.161
		choose	22	18	-0.079
		chosen	3	4	-0.273
		chose	5	6	0.04
		alternative	0	1	NA
		option	13	13	-0.632
		decision	7	5	-0.248
		prefer	6	8	0.402
		options	13	13	-0.632
		allow	51	34	0.345
		I see	0	0	NA
		agree	34	18	0.062
		any questions	20	0	-0.171
		brave	1	4	0.646
		can you	393	0	-0.572
		either	56	36	-1.334
		feel free	1	0	NA
		flexible	1	2	0.646
		free	17	18	0.652
		freedom	2	2	-0.065
		guide	3	3	1.119
		independently	5	3	0.321

	initiative	6	2	1.583
	opportunity	11	9	0.514
	possible	46	12	-2.339
	decide	11	14	0.327
	interesting	28	15	0.946
	create	35	7	1.16
Competence Supportive				
	win	17	18	0.8
	high score	3	0	-0.313
	highest	63	10	-0.411
	fastest	4	3	0.226
	winning	17	18	0.8
	reward	1	2	0.646
	wins	17	18	0.8
	terrific	0	0	NA
	correct	84	29	-0.643
	very good	54	0	0.518
	smart	9	12	0.182
	better	65	50	0.147
	top	70	37	0.076
	award	6	6	-0.386
	best	32	34	0.306
	champ	10	3	2.044
	clever	5	8	1.445
	excellent	119	34	0.99
	fabulous	1	1	0.646
	good	938	66	0.503
	great	69	37	0.732
	nice	222	48	0.864
	perfect	40	26	-0.927
	proud	2	3	0.914
	capable	4	4	-0.093
	good job	62	0	0.335

fantastic	36	14	1.288
try hard	0	0	NA
can do	95	0	1.028
better than	7	0	-0.824
ability	4	7	-0.093
able	57	36	-0.135
accomplishment	0	0	NA
achievable	7	3	1.71
admirable	0	0	NA
awesome	27	18	0.795
bonus	4	2	1.293
brilliant	5	5	0.04
can	1650	66	1.211
confident	8	7	0.319
congratulate	5	5	0.321
demonstrate	3	6	0.105
doing well	8	0	0.248
effort	11	17	0.137
example	38	24	-0.394
feedback	2	2	0.914
go on	66	0	-0.02
goal	39	26	0.175
good boy	18	0	0.729
good girl	34	0	0.458
help	169	55	0.611
hint	4	2	1.293
intelligent	0	0	NA
learn	53	32	0.238
nice one	2	0	-0.1
practice	63	34	0.877
praise	0	0	NA
role model	0	0	NA
strength	0	2	NA

	strong	8	11	0.097
	successful	0	3	NA
	superb	0	0	NA
	technique	22	18	0.058
	tips	8	12	-0.131
	tremendous	0	0	NA
	well done	272	0	-1.326
	wise	1	1	0.646
	wonderful	13	11	-0.823
	work hard	1	0	NA
	try	290	63	0.949
	amazing	2	4	0.914
	wow	22	19	0.058
	I like	0	0	NA
	challenge	26	14	0.268
	improve	7	11	0.24
	revise	15	10	-0.265
	guidance	0	0	NA
	coaching	3	2	-0.273
	instruction	30	29	-0.741
	great work	7	0	0.173
	innovative	0	0	NA
	nicely done	0	0	NA
Relatedness Supportive				
	apology	2	2	0.914
	assist	3	3	0.105
	fair	10	19	0.057
	we	1737	66	2.204
	our	261	58	2.412
	care	50	32	-0.799
	cared	50	32	-0.799
	caring	50	32	-0.799
	together	102	47	0.801

		each other	43	0	0.924
		share	11	11	0.696
		we can	138	0	0.474
		accept	16	10	-0.023
		all	1750	66	-1.337
		belong	0	0	NA
		buddy	7	14	-0.248
		classmate	2	2	NA
		cooperation	53	13	0.587
		empathy	0	0	NA
		encourage	3	2	NA
		everyone	236	61	0.921
		friendly	0	0	NA
		group	159	35	1.952
		group work	1	0	NA
		look after	1	0	0.622
		mate	126	34	1.019
		peers	3	5	1.119
		sorry	218	60	-0.823
		team	313	41	0.476
		teammate	12	13	-0.161
		teamwork	7	5	0.473
		welcome	15	12	-0.096
		work together	11	0	0.61
		friends	10	16	0.258
		connected	4	4	-0.093
		like me	2	0	0.88
Need Thwarting	Autonomy Thwarting				
		punish	1	4	NA
		stop	230	59	1.408
		penalty	2	2	NA
		rule	65	36	0.436
		stop it	11	0	0.354

	shut up	0	0	NA
	boring	8	8	0.138
	cannot	15	20	0.618
	chaos	0	0	NA
	childish	0	0	NA
	deny	0	0	NA
	don't	673	66	0.813
	forbidden	0	0	NA
	force	6	2	-0.454
	has to	29	0	0.473
	have to	316	0	1.124
	hurry	60	31	0.64
	immediately	1	1	0.742
	keep working	7	0	-0.195
	limit	4	3	0.097
	must	22	26	-0.062
	never	25	25	0.711
	prohibit	0	0	NA
	should	211	54	0.453
	Shouldn't	0	0	NA
	got to	127	0	0.461
	no	951	66	1.571
	need to	608	0	-0.032
	pay attention	20	0	0.937
	shh	222	25	1.897
	deadline	0	0	NA
	in charge	7	0	-0.471
	command	5	2	NA
	conform	0	0	NA
	told you	10	0	0.866
Competence Thwarting				
	lose	10	20	0.366
	worst	3	4	NA

	mistake	25	16	0.039
	dropout	0	0	NA
	worse	4	7	NA
	awful	7	13	0
	bad	43	32	0.78
	wrong	52	39	0.258
	idiot	0	1	NA
	silly	23	24	0.305
	stupid	0	0	NA
	suck	0	1	NA
	keep up	2	0	1.087
	bottom	42	24	0.212
	clumsy	0	0	NA
	dumb	2	1	NA
	fail	5	3	0.259
	failure	0	0	NA
	give up	5	0	-0.304
	talented	1	1	NA
	unable	1	1	NA
	luck	26	13	0.899
	disorganized	0	0	NA
Relatedness Thwarting				
	disrespectful	2	2	NA
	angry	2	2	NA
	annoy	5	9	0.259
	blame	2	3	1.049
	crap	1	2	0.742
	crazy	5	8	-0.043
	cruel	0	1	NA
	cry	1	2	0.742
	disappointed	4	3	-0.25
	dislike	0	0	NA
	embarrassing	3	4	-0.114

fool	2	3	0.069
foolish	0	0	NA
horrible	4	2	1.484
immature	0	0	NA
isolated	1	1	0.742
jealous	1	1	NA
nasty	0	1	NA
neglect	0	0	NA
rude	7	6	0.256
shame	2	1	1.049
ugly	0	1	NA
weird	8	9	-0.104
sick of	4	0	1.538

Appendix C (Chapter 4) – Classification of TMBs Study Supplementary Material

Appendix C.1

Delphi Round 1 Results with Plots

TMB#1

Provide Rationales

Description:

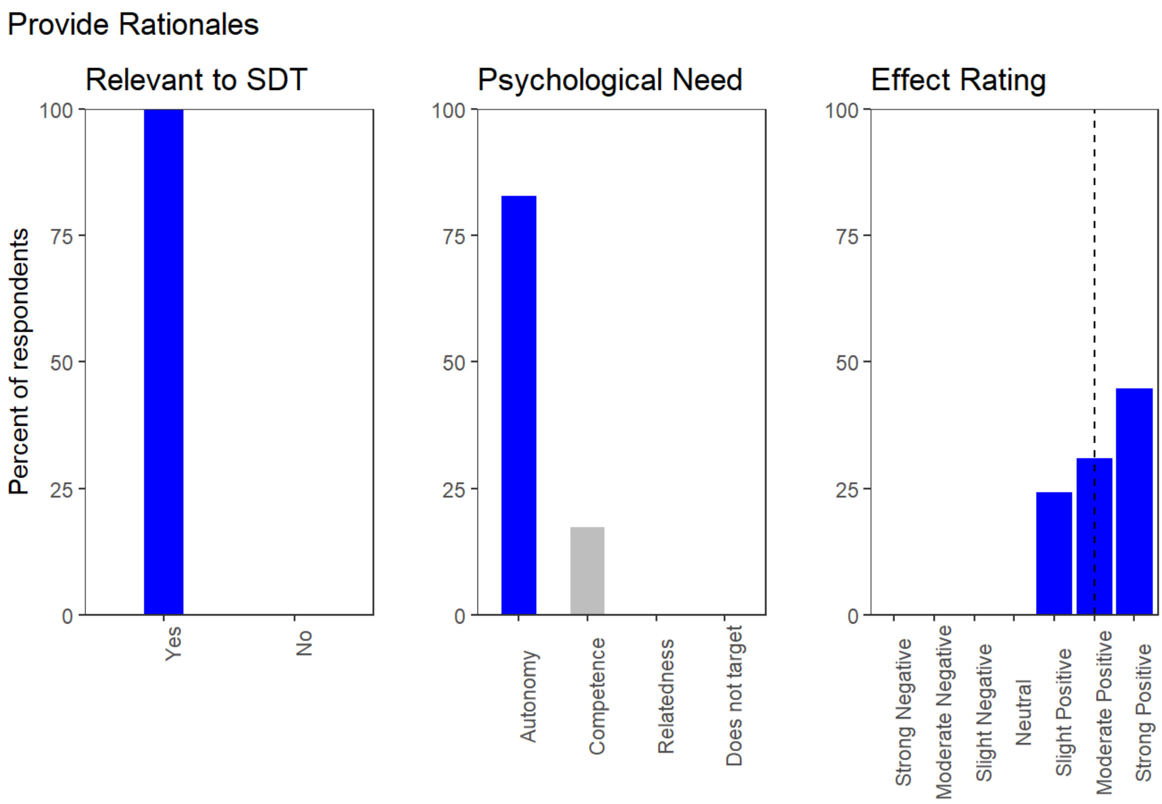
Explain why an activity is important, or how it might be useful

Example Behaviour:

"Being active and doing more exercises will help you to have healthier body"

Function Description:

Students understand the benefit of the task to them and their lives



TMB#2

Student input or choice

Description:

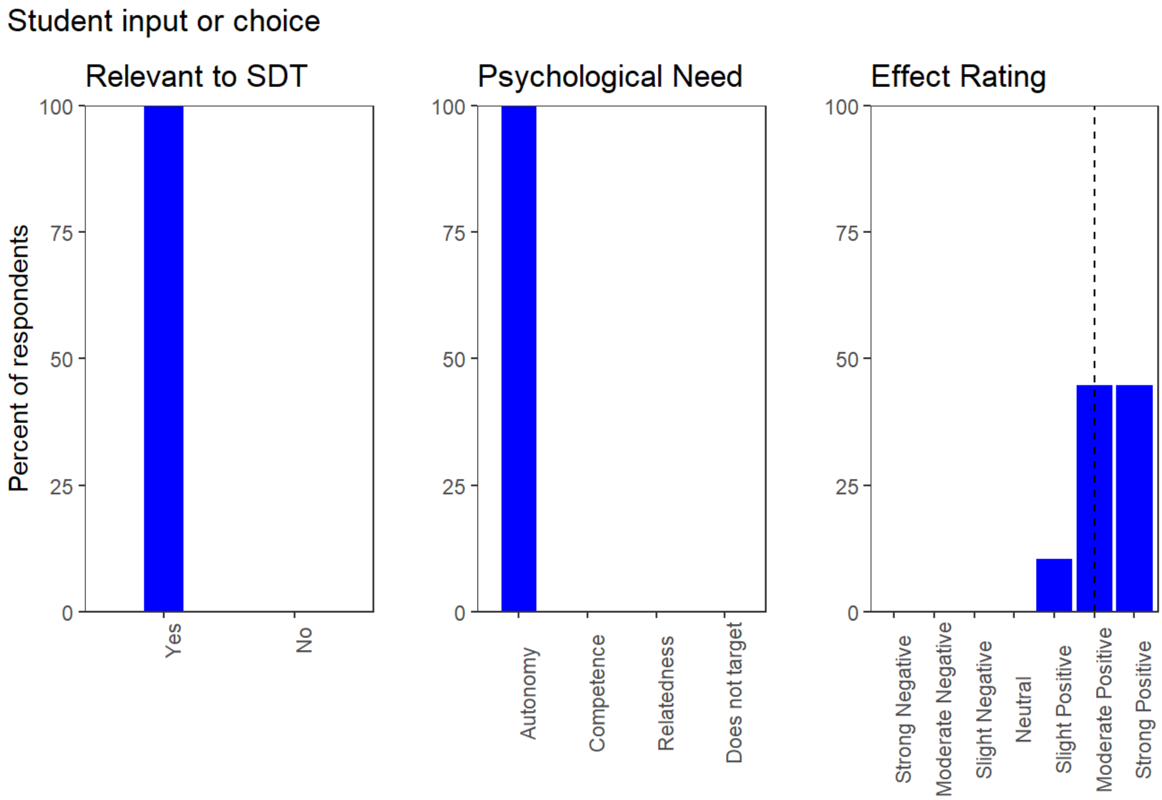
Let students have some input or make choices about the things they do in class

Example Behaviour:

"You can either work with a friend or do it by yourself"

Function Description:

Allows students to choose tasks that align with their priorities and capabilities



TMB#3

Active Learning

Description:

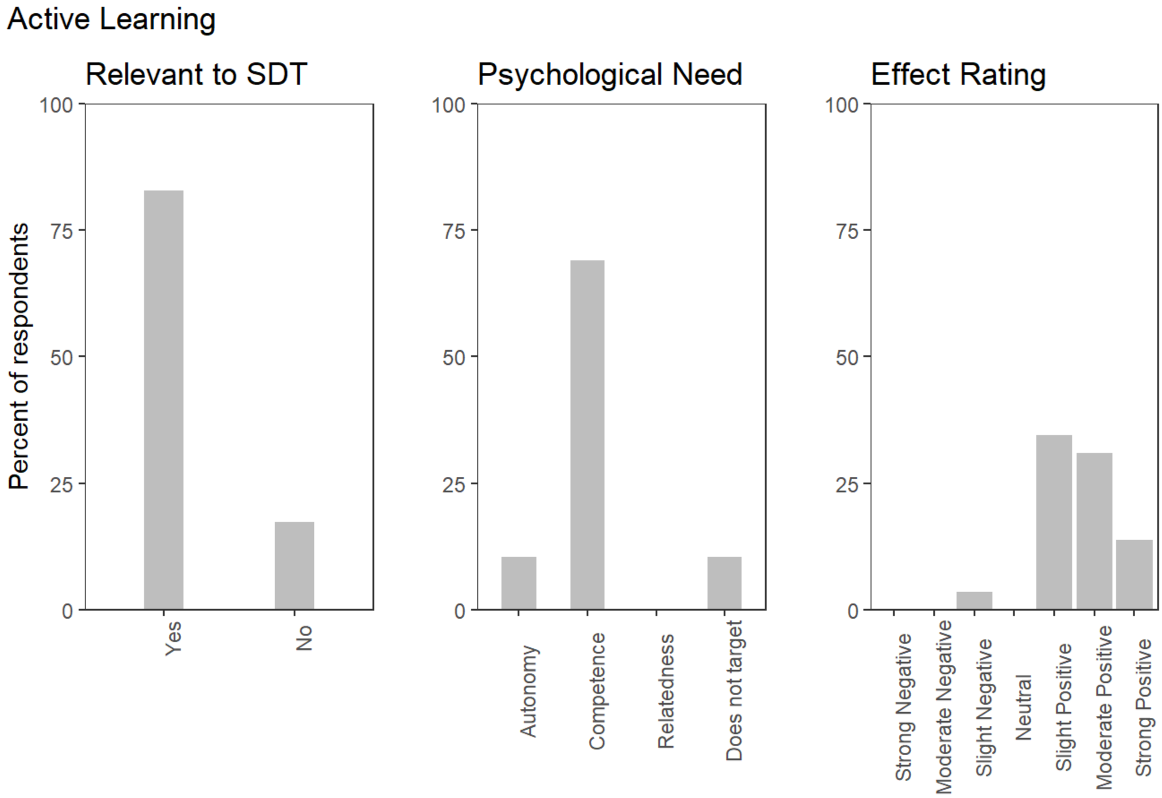
Set up activities where all students are busy doing a task or solving a problem

Example Behaviour:

"Use these numbers to figure out how heavy the Sydney Harbour Bridge is"

Function Description:

Allows hands-on practice with a skill to develop a sense of progress and mastery



***

TMB#4

Promoting cooperation

Description:

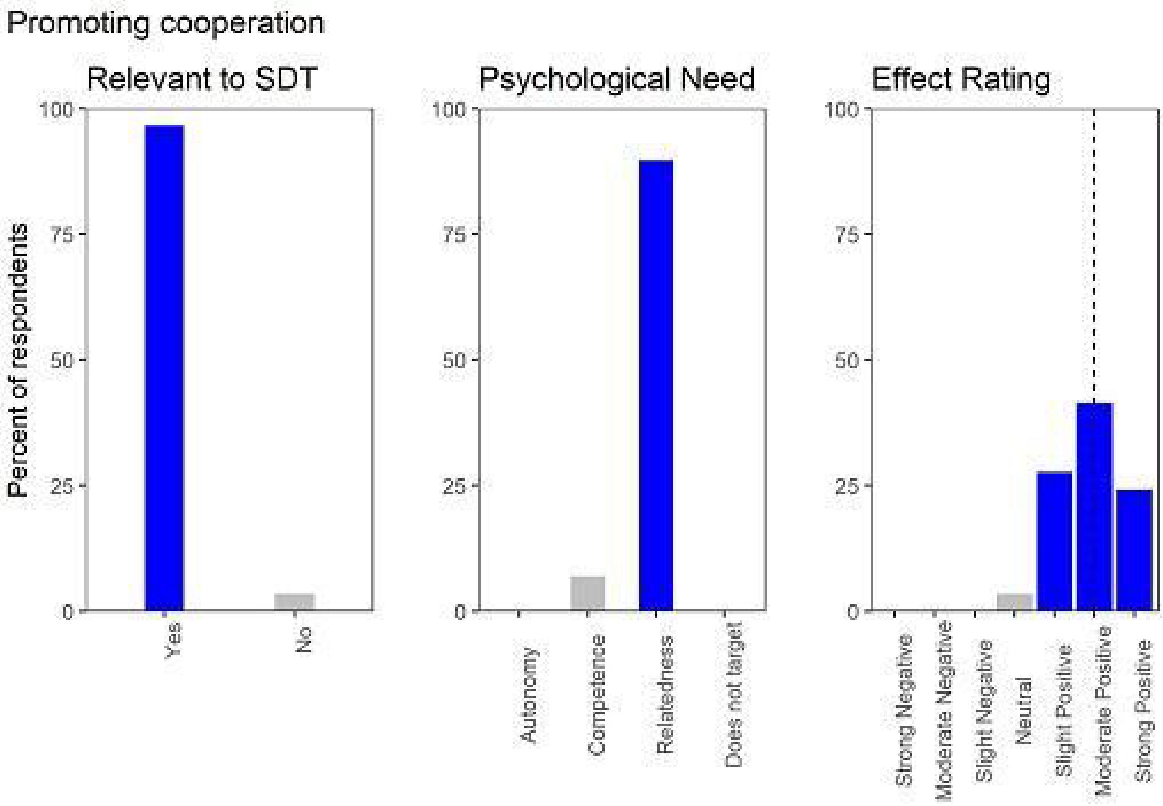
Set up activities that let students work together on tasks

Example Behaviour:

"As a group, try to figure out this answer"

Function Description:

Allows joint pursuit toward a goal and peer feedback on progress



***

TMB#5

Modelling correct technique

Description:

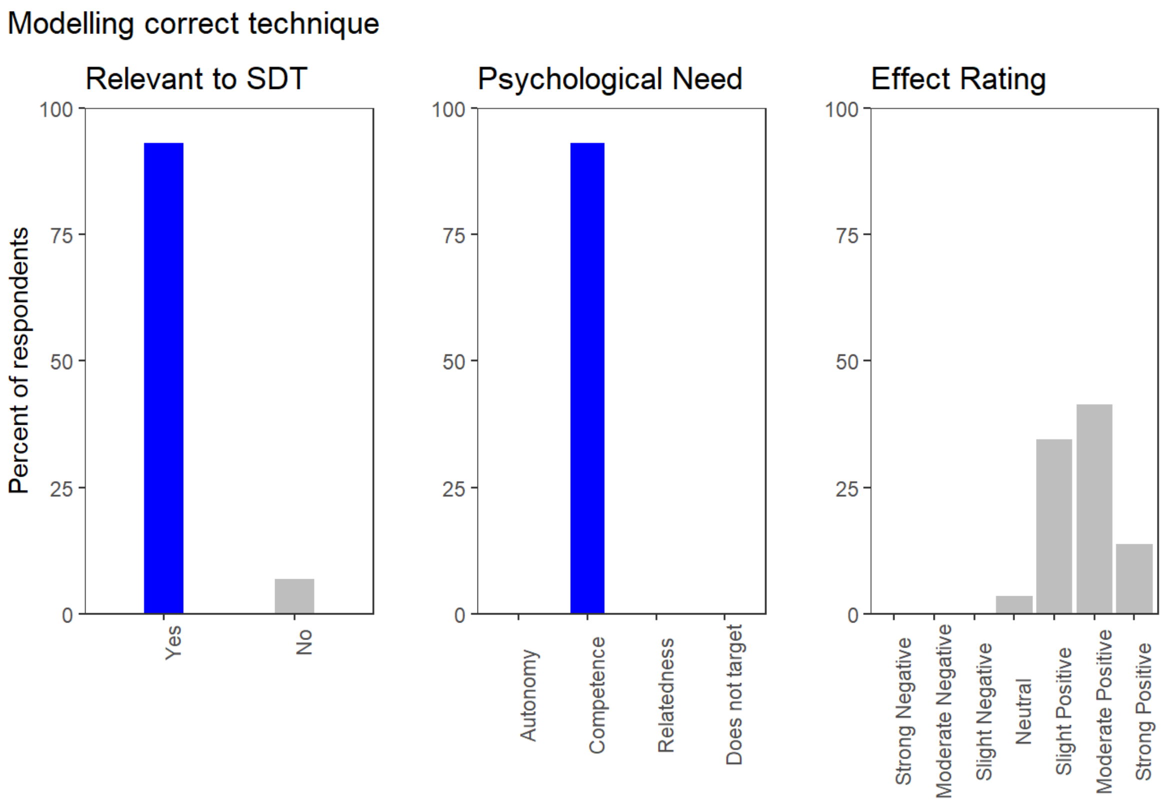
Modelling or demonstrating correct technique

Example Behaviour:

PE - "When throwing, see how my other hand points at the target?"; Maths - "Notice how if we carry the one, it helps with the next step"

Function Description:

Makes skill look achievable and provides template for student to follow



***

TMB#6

Differentiating and scaffolding

Description:

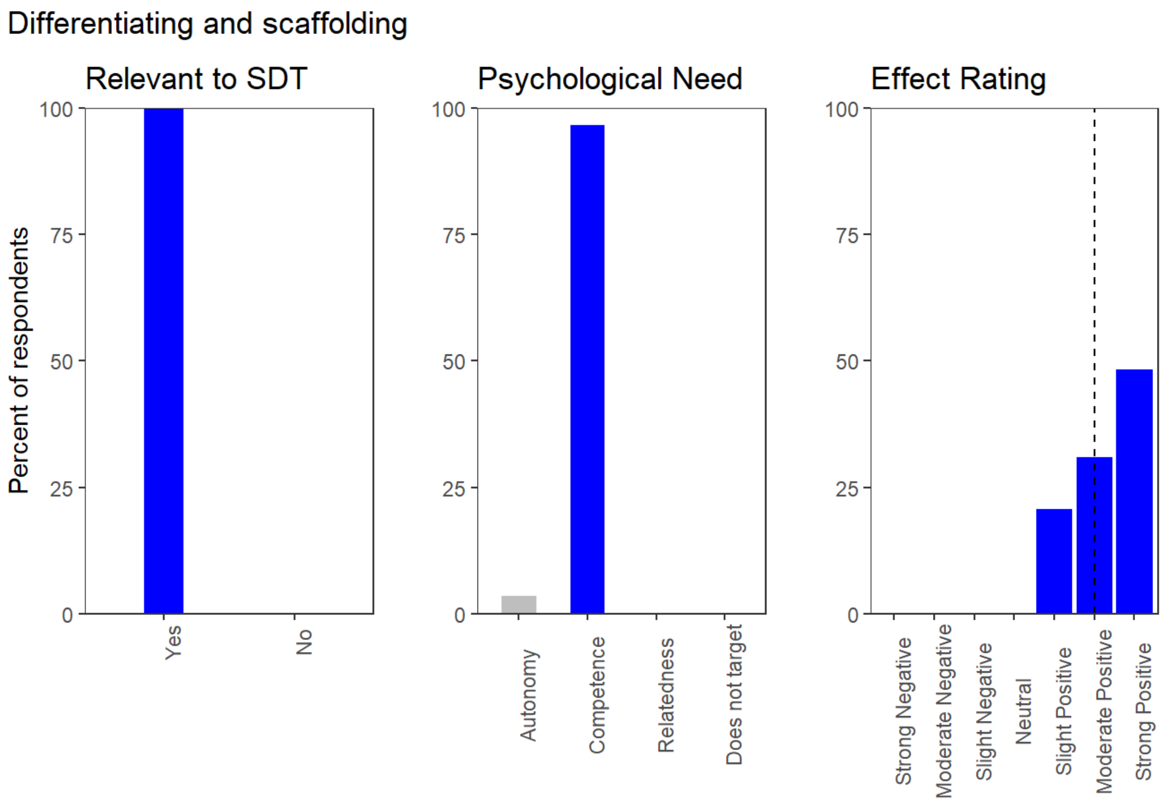
Give students harder tasks if they find it too easy, or easier tasks if they find it too hard

Example Behaviour:

"Start with question 6, and if it is too hard, go back to question 1, or if you finish quickly, try question 13"

Function Description:

Students get the right amount of challenge for them



***

TMB#7

Humour

Description:

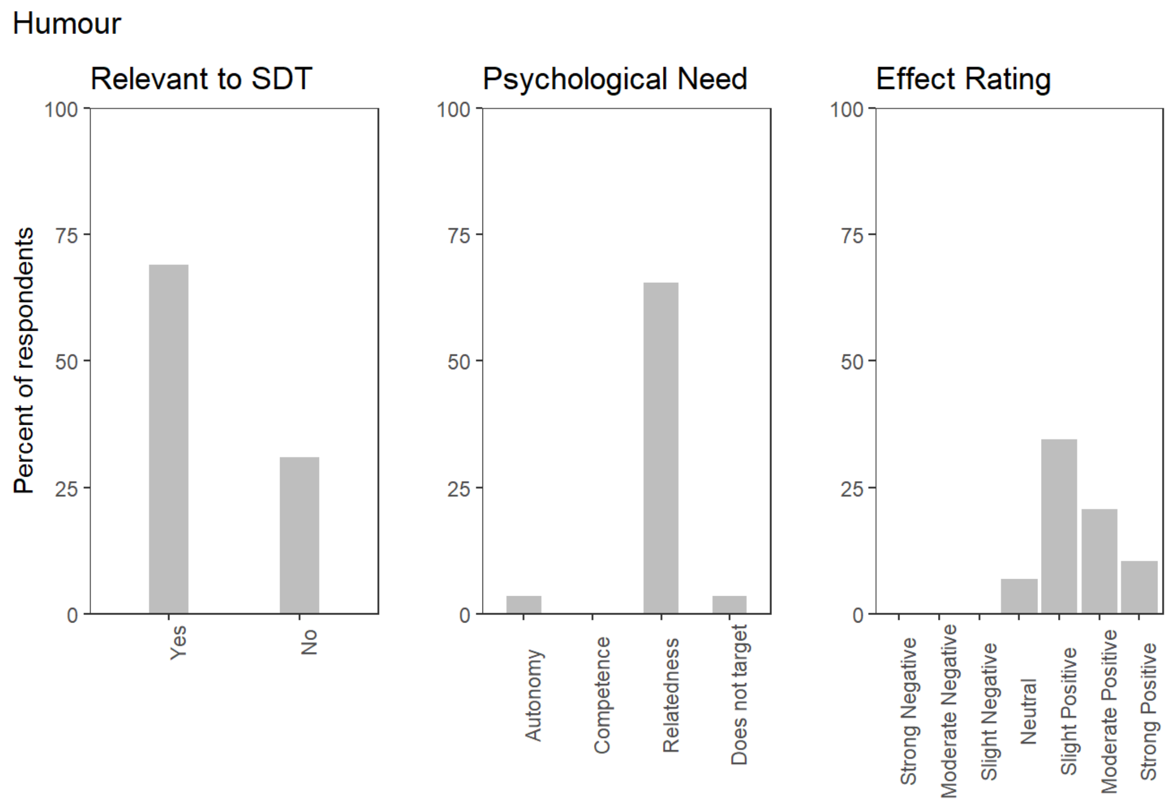
Use humour so the class is fun

Example Behaviour:

"What did the triangle say to the circle? You are pointless"

Function Description:

Alleviates anxiety and reduces goal-focus; increases warmth for teacher



***

TMB#8

Attending to students

Description:

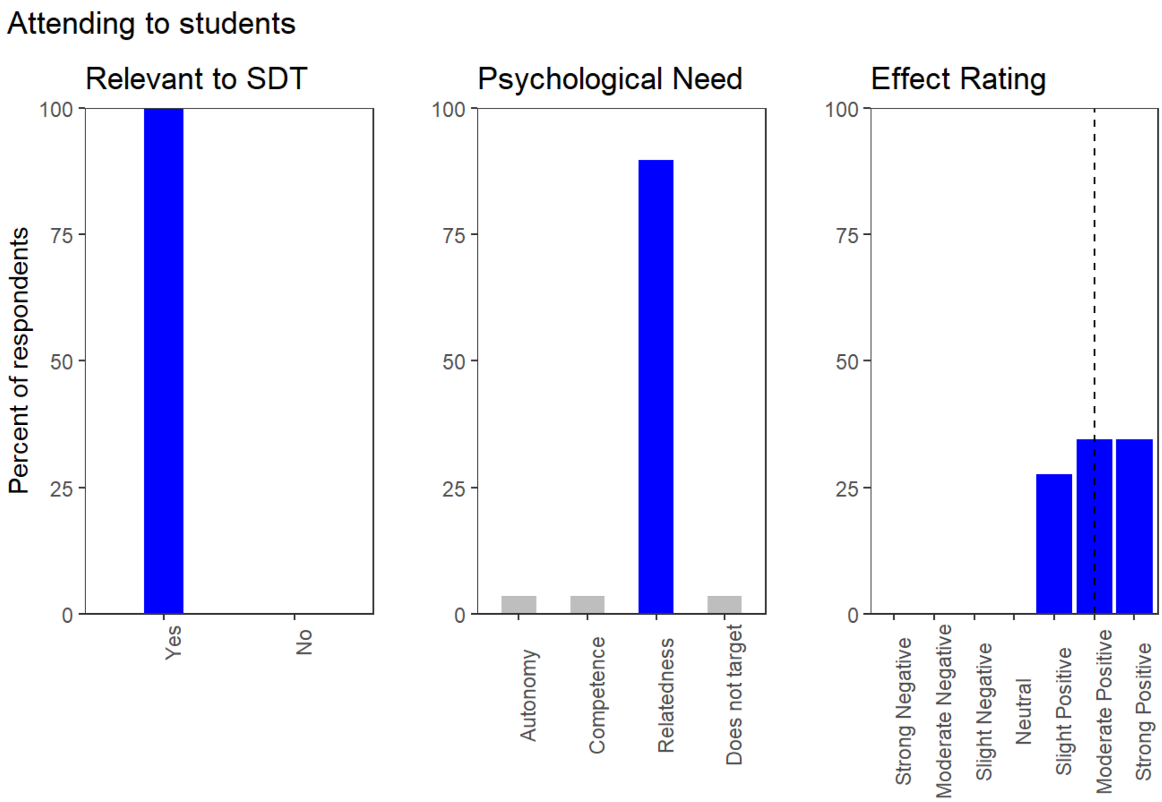
Pay close attention to students while they are in class

Example Behaviour:

Approach students and pay attention to students while they are performing

Function Description:

Makes students feel valued or cared for and that their efforts are noticed



***

TMB#9

Unconditional positive regard

Description:

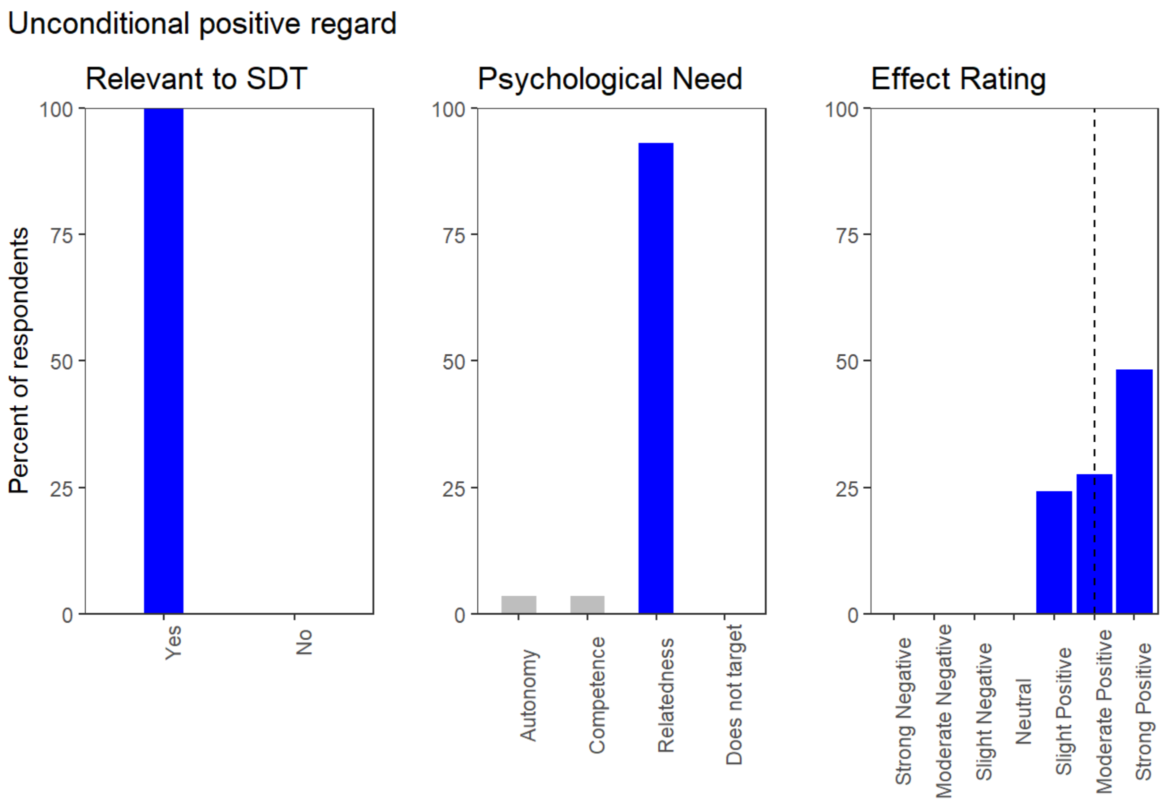
Act warmly towards students even ones who are challenging

Example Behaviour:

The teacher is kind even to one student who did a task incorrectly and another who did not complete the task.

Function Description:

Ensures performance mistakes or behavioural misconduct are not met with ego-threatening behaviour



***

TMB#10

Ask about students progress, welfare, and feelings

Description:

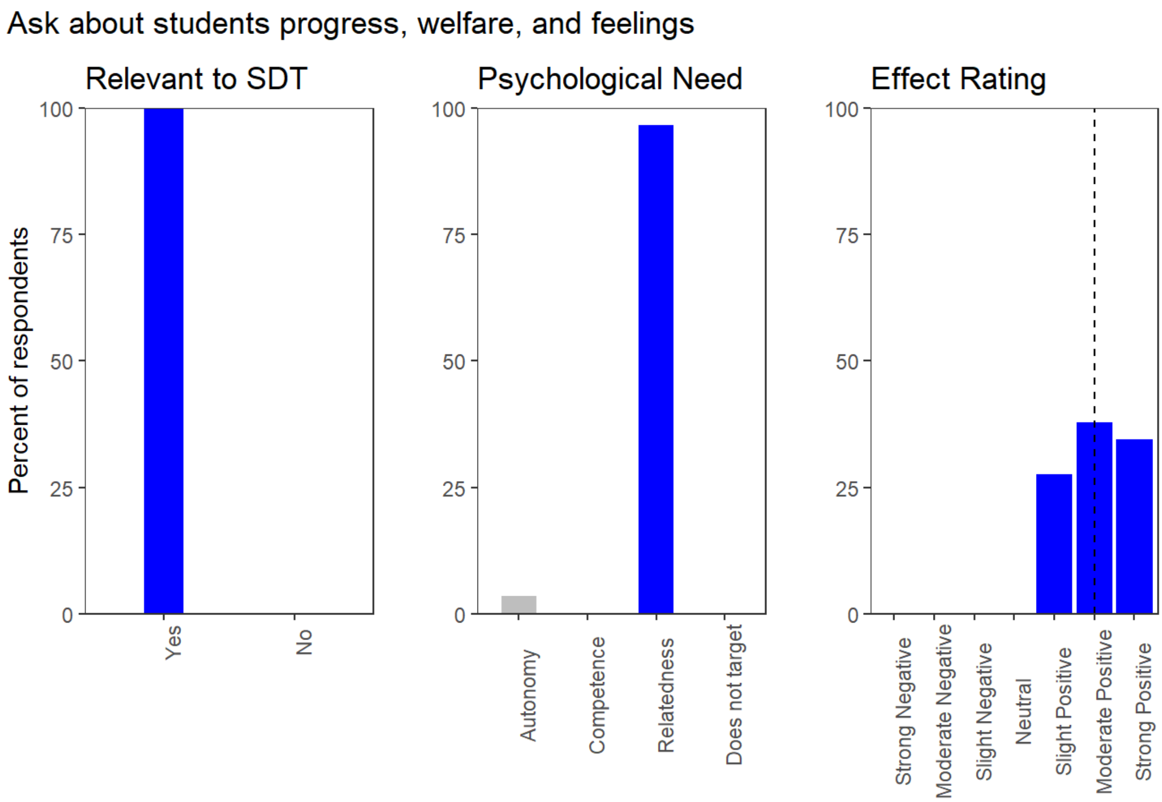
Enquiring about students progress, welfare or feelings

Example Behaviour:

"How was your dance recital on the weekend, John?"

Function Description:

Shows care and encourages students to express themselves openly to connect with their teacher



***

TMB#11

Discuss values

Description:

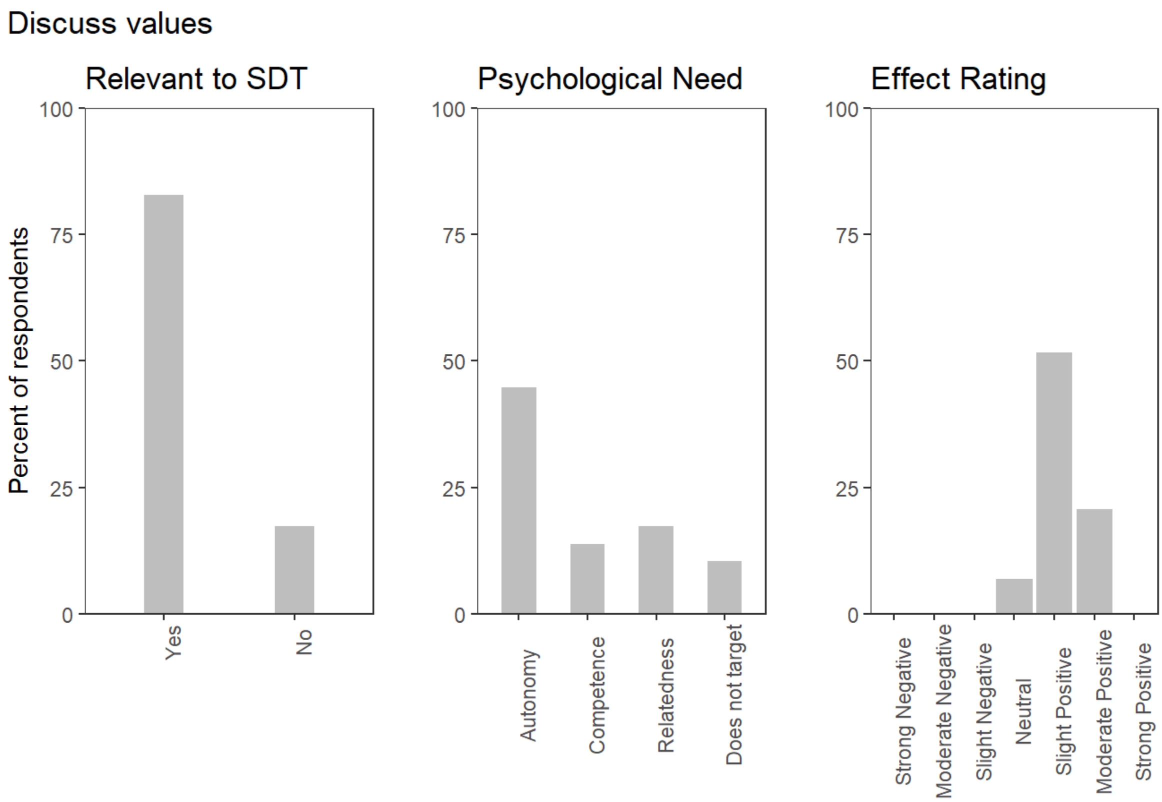
Discusses values expectated in the class. Does not include offer of rewards or punishments.

Example Behaviour:

"In this class, what do you think our values should be?"

Function Description:

Makes explicit the personal characteristics being developed by the class environment.



***

TMB#12

Hope, encouragement, optimism

Description:

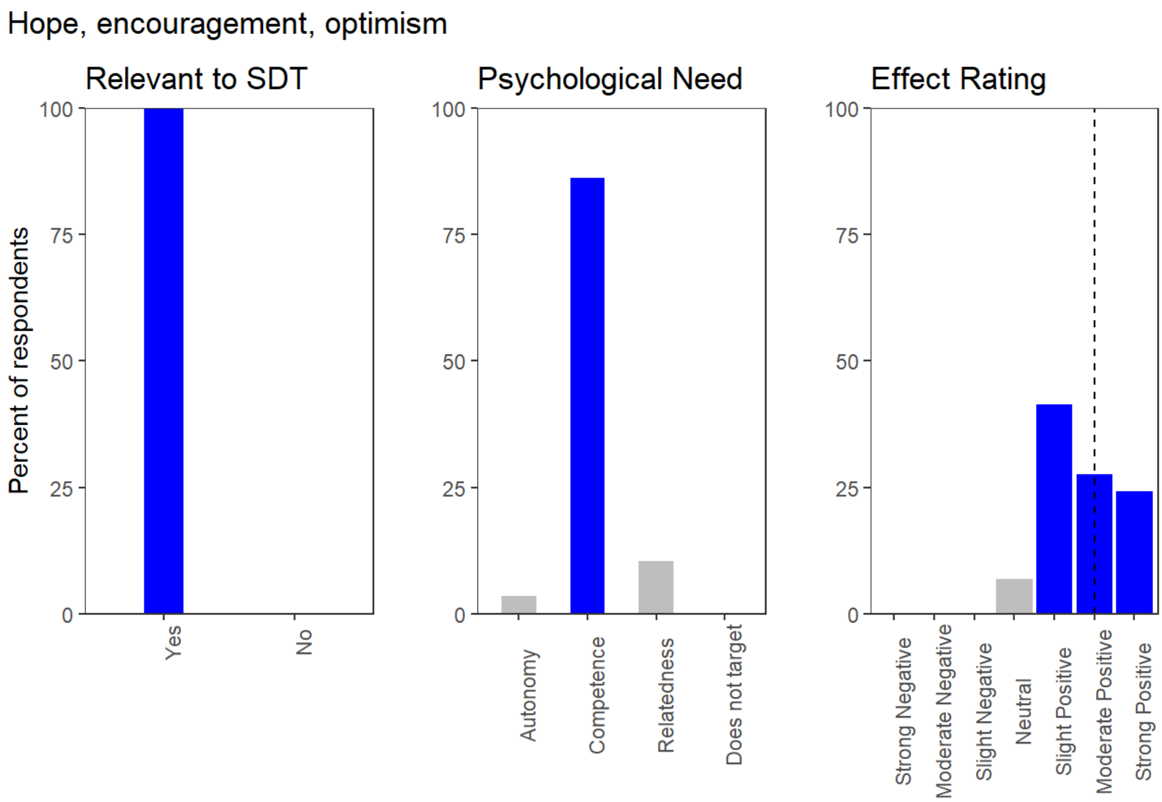
Expressing hope and optimism in students potential to succeed in the future

Example Behaviour:

"I know you can do this"

Function Description:

Improves perceived ability to meet goals



***

TMB#13

Showing vulnerability or humility

Description:

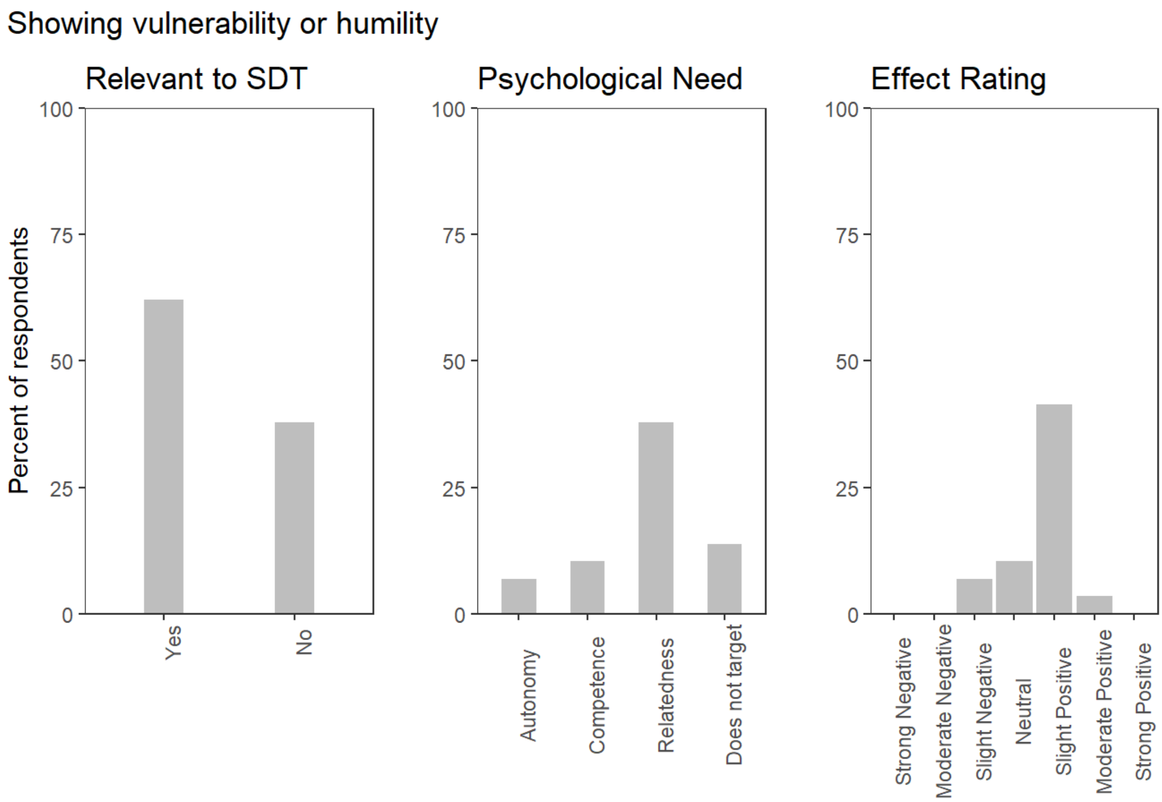
Actions to reduce the status of the teacher (e.g., via self-deprecation, acknowledgment of fallibility or difficulties now or in the past)

Example Behaviour:

"My little chicken arms"

Function Description:

Reduces the pressure students might perceive, which inherently comes from the teachers authority and status



***

TMB#14

Expressing Affection

Description:

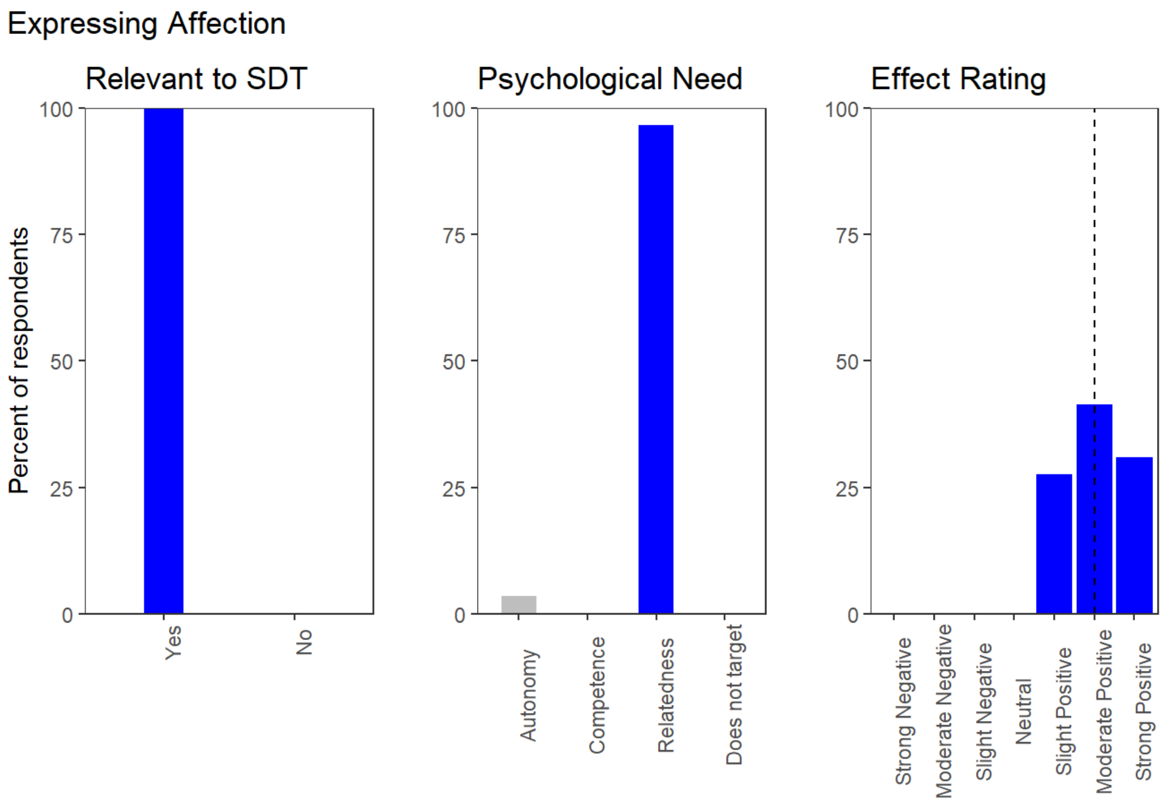
Be warm and kind to students

Example Behaviour:

"It is good to see you, Theresa!"

Function Description:

Students feel they are cared for



***

TMB#15

Teacher Enthusiasm

Description:

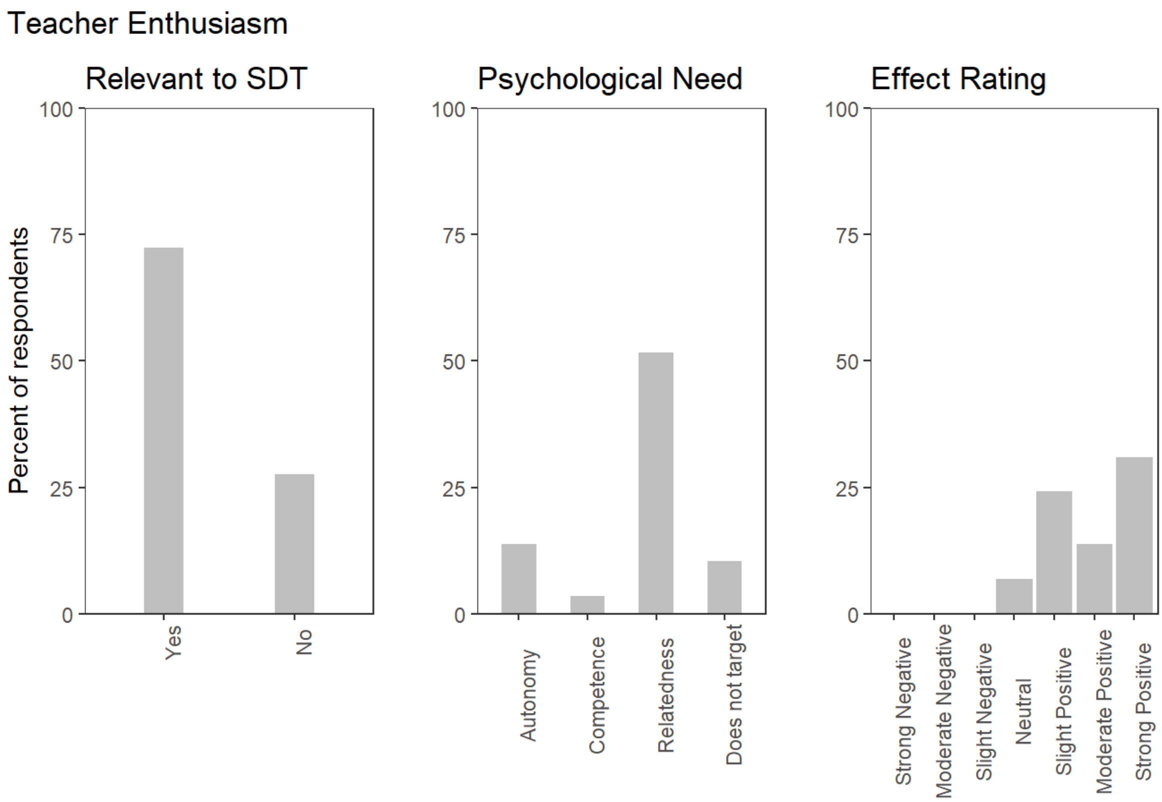
Present content enthusiastically to make things fun and interesting

Example Behaviour:

"All rightey then! Let us get ready to ruuuummmblllleeee!"

Function Description:

Models the attitude and energy that the teacher would like the students to experience.



***

TMB#16

Outlining Reward Contingencies

Description:

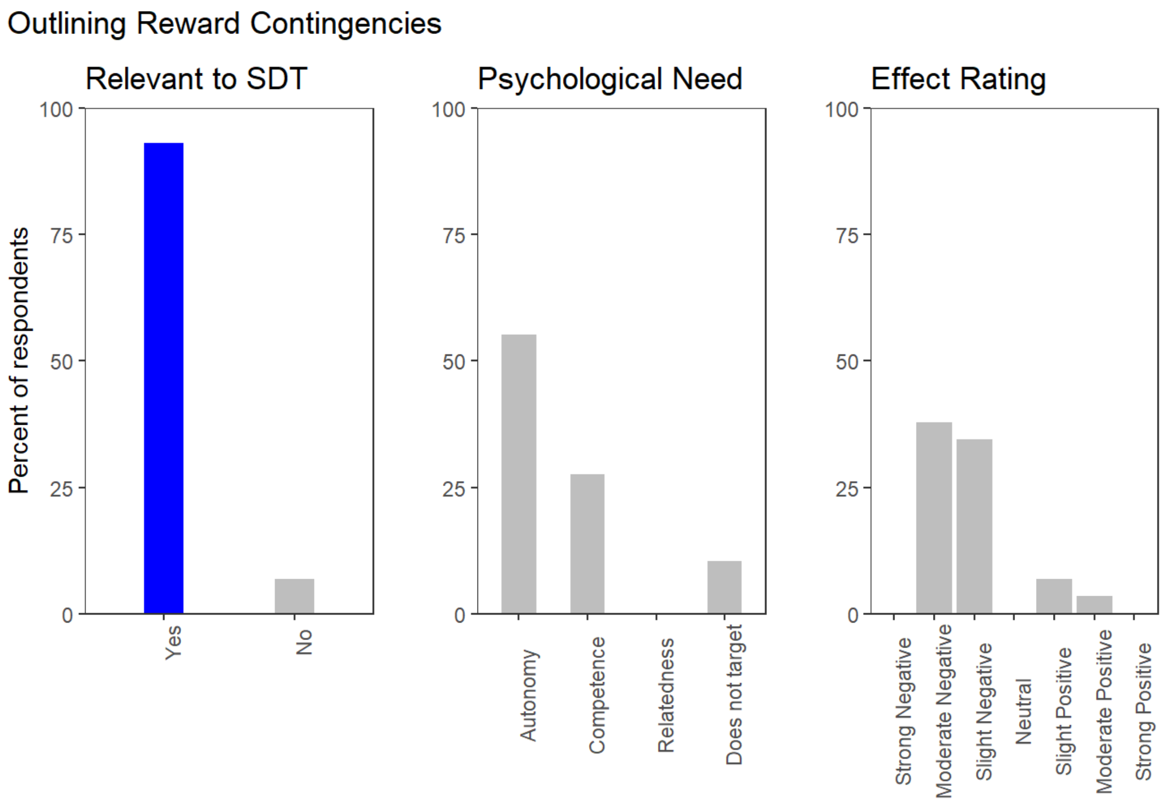
Offering (but not yet providing) if-then extrinsic rewards—privileges/items that are not inherent to the task and are provided in an effort to promote a behaviour

Example Behaviour:

"If you work hard, you can finish early"

Function Description:

To direct and structure behaviour so students know what behaviour is valued



***

TMB#17

Fair use of praise and rewards

Description:

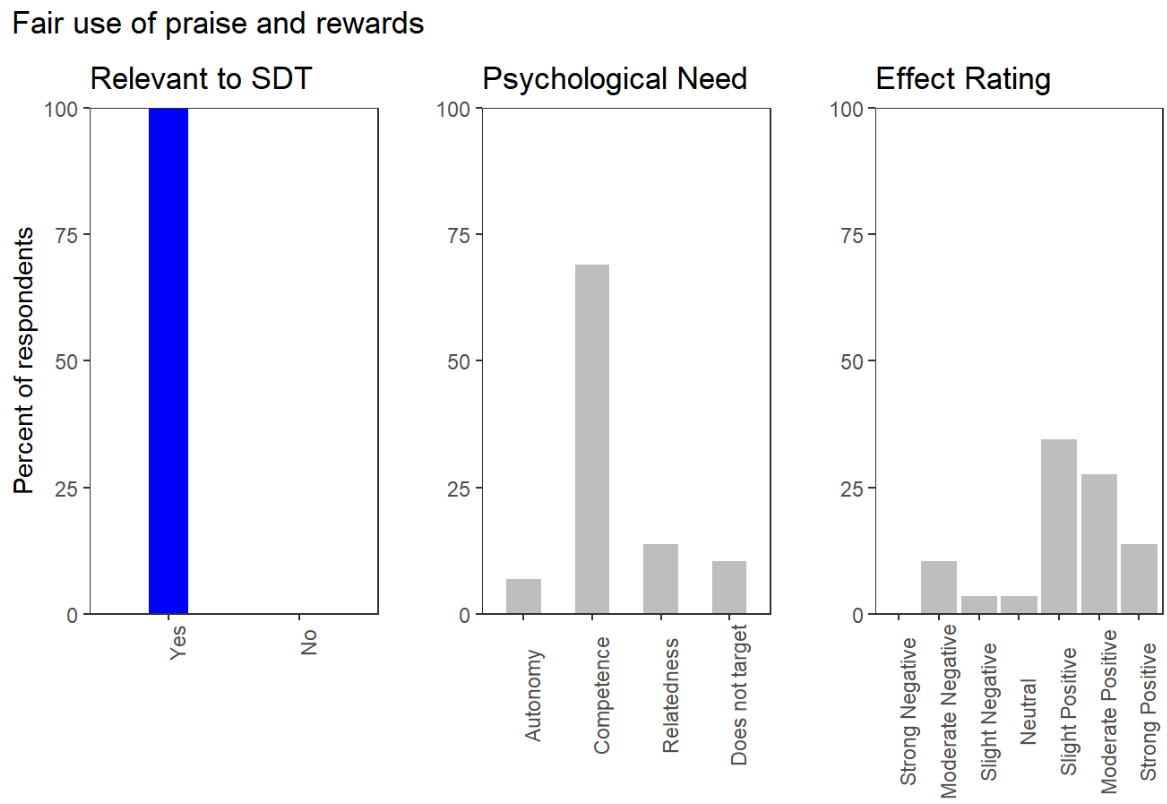
Praises and rewards students fairly and equally; does not show favorites

Example Behaviour:

Complementing all three people who completed a project in specific ways

Function Description:

Increases sense of efficacy, inclusion and belonging



***

TMB#18

Feedback aimed at improvement or effort

Description:

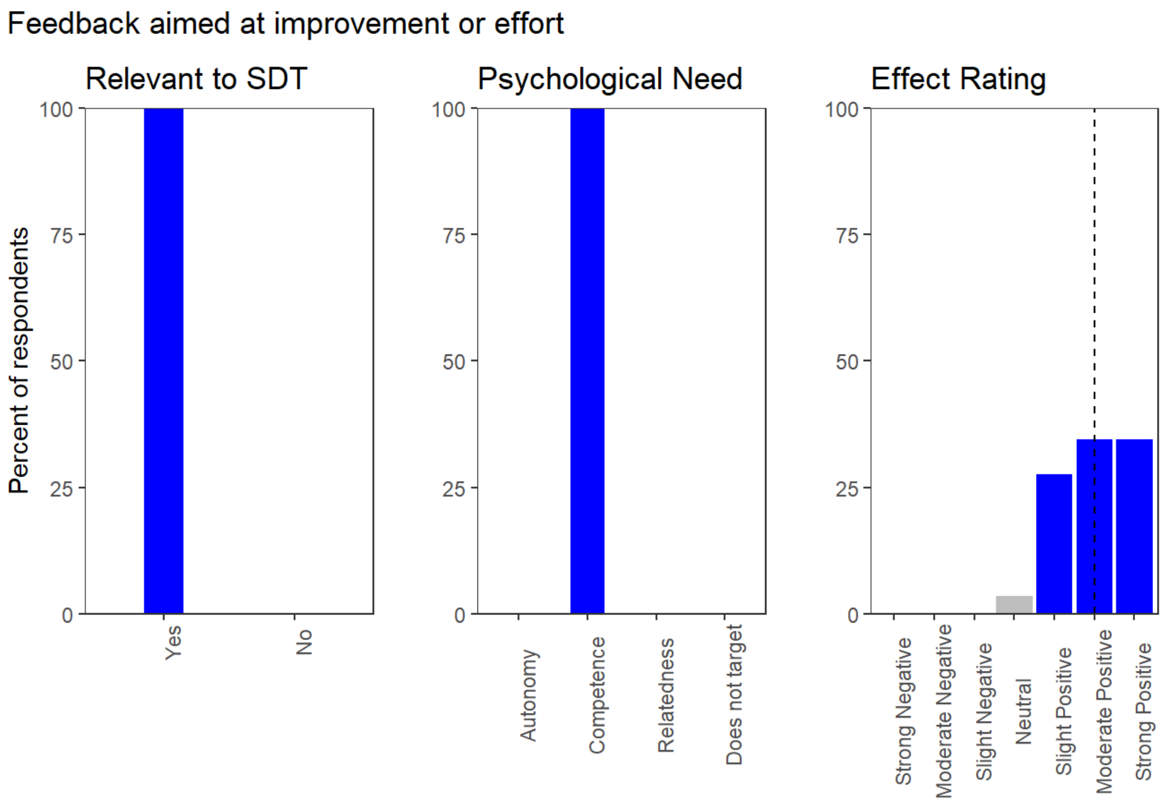
Provides critical feedback to help a student improvement or increase effort

Example Behaviour:

"If you combine these two rules, it will help get that solution."

Function Description:

Emphasises the malleability of abilities critical for success.



***

TMB#19

Feedback - Private

Description:

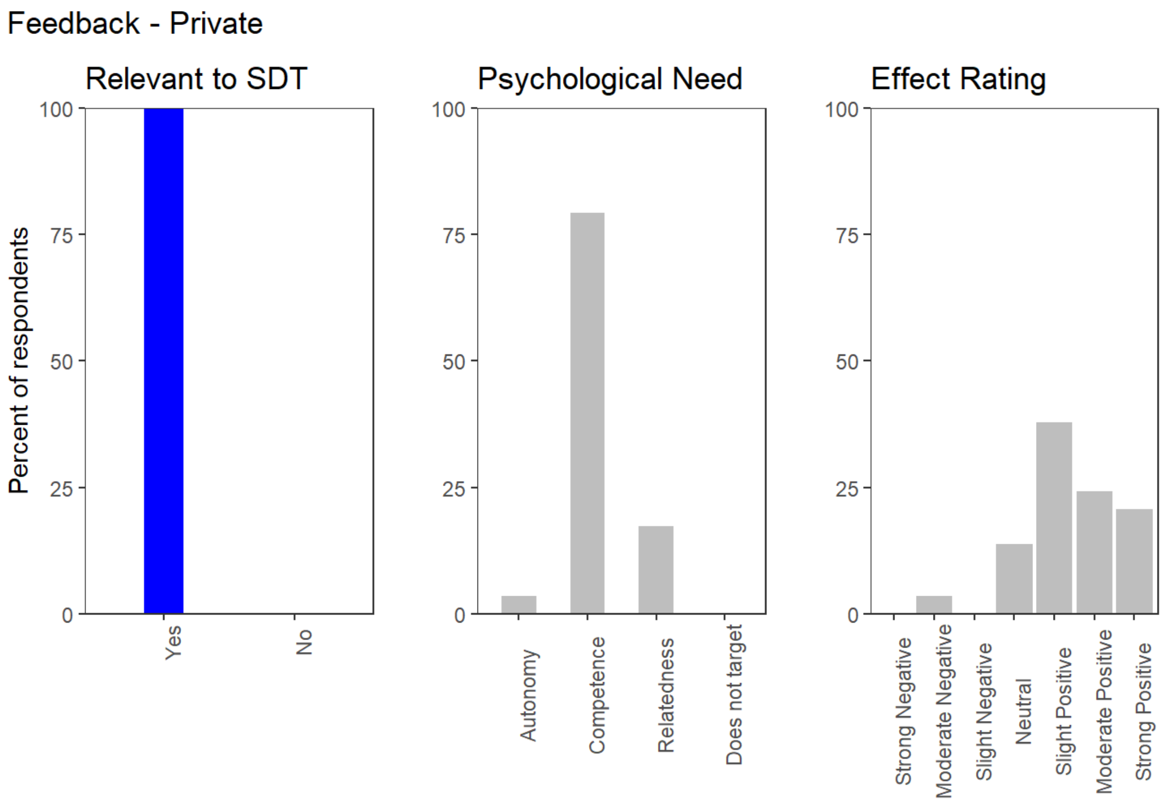
Provide any kind of sensitive or critical feedback in private

Example Behaviour:

Provide feedback 1 on 1 with the student

Function Description:

Mitigates risk of feedback being ego-threatening



***

TMB#20

Feedback - Specific

Description:

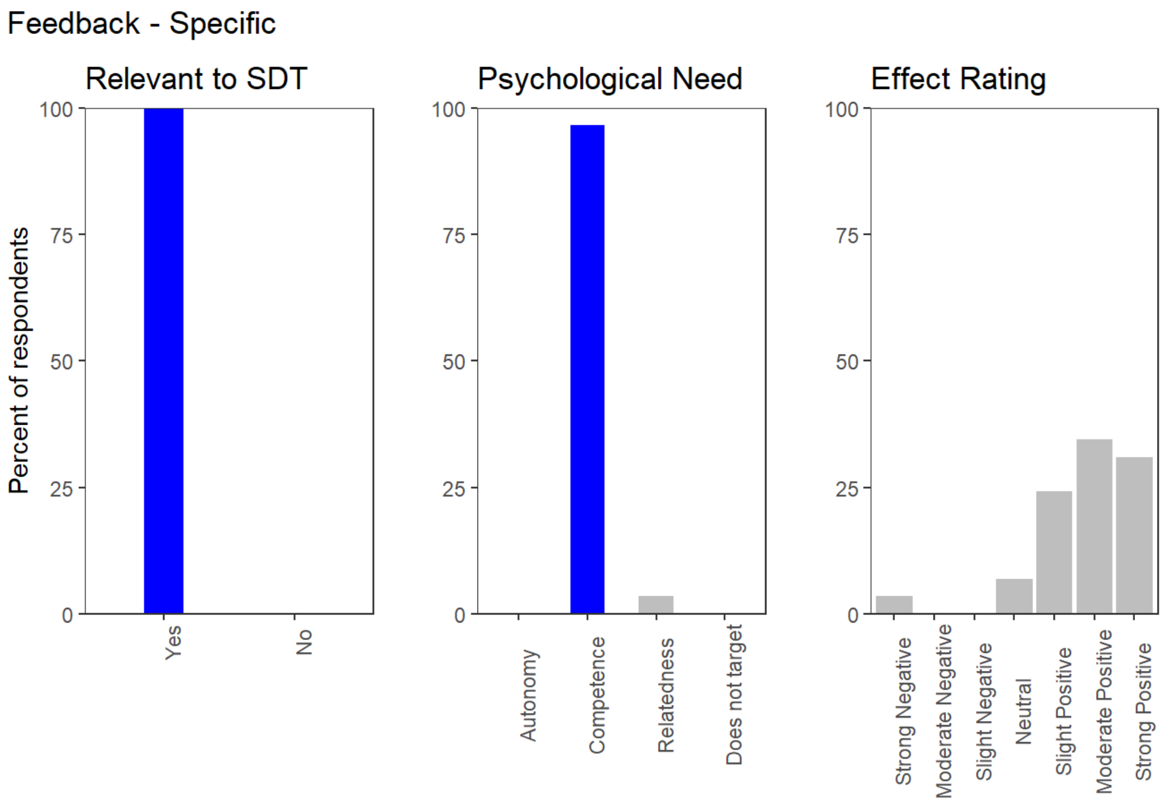
Provides feedback that specifically targets a defecit or strategy for improvement

Example Behaviour:

"The reach of your arms is not long enough for interceptions", "If you keep your eye on your attacker then you can try for an intercept, but focus on marking your girl"

Function Description:

Clarifies path toward goal achievement.



***

TMB#21

Frequent criticism

Description:

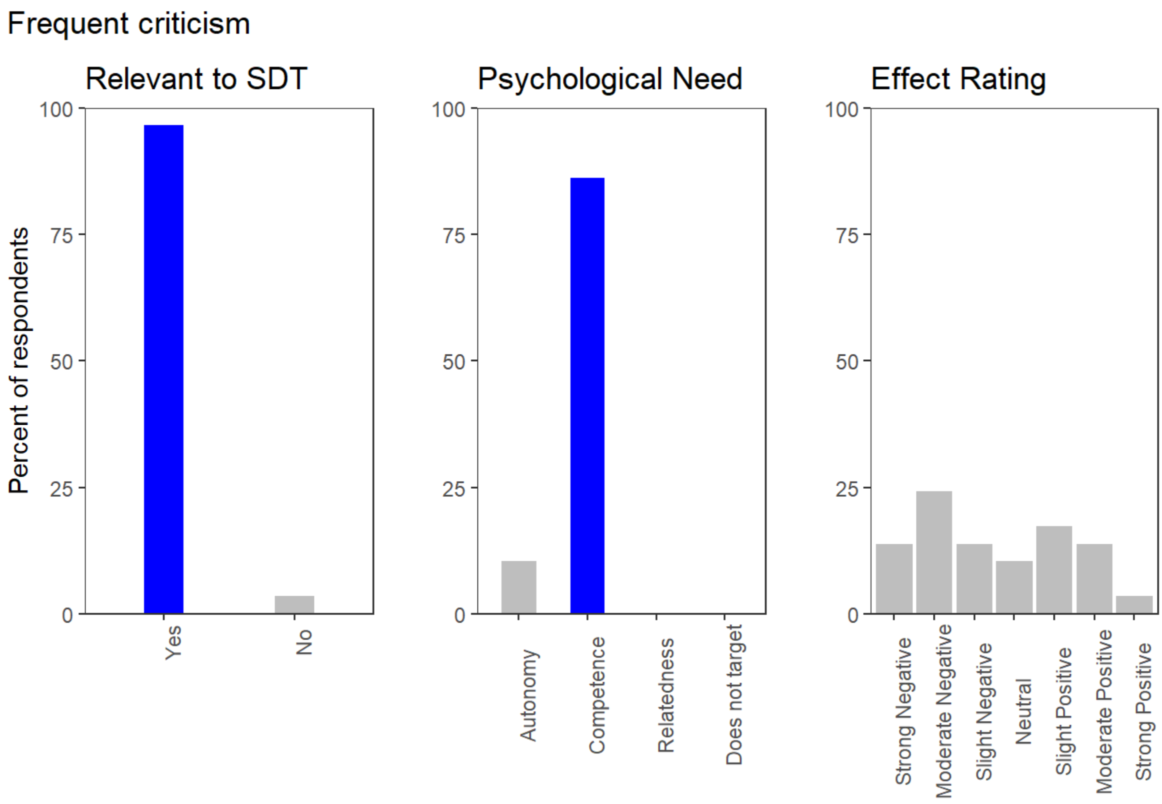
Frequently provide critical feedback to students

Example Behaviour:

Consistently patrolling the class with suggestions for getting unstuck

Function Description:

Promotes continual improvement in abilities.



***

TMB#22

Offering hints

Description:

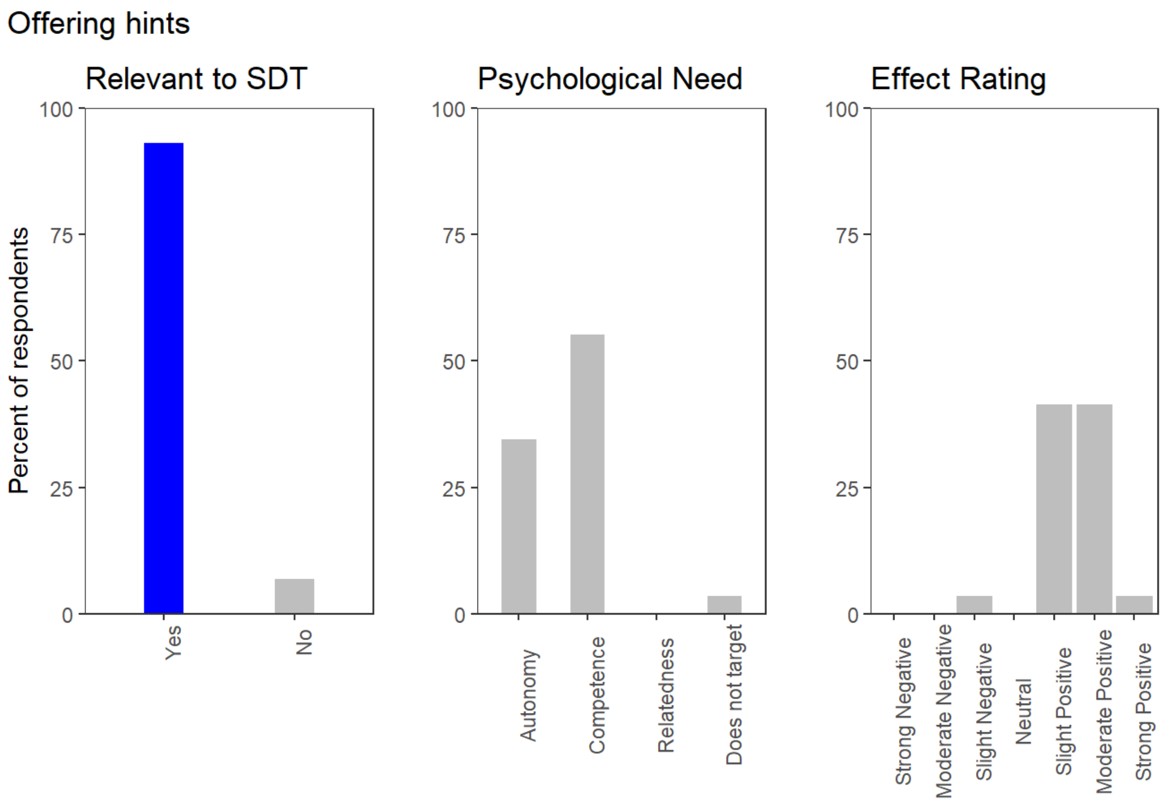
Give hints to help students along without giving them the "right answer"

Example Behaviour:

“Putting the number on your calculator seems to work better than keeping it in your head”; “It might be easier to start with this formula.”

Function Description:

Supports the student’s own learning processes. Allows students to maintain an internal locus of causality during learning.



***

TMB#23

Praise / fixed-quality

Description:

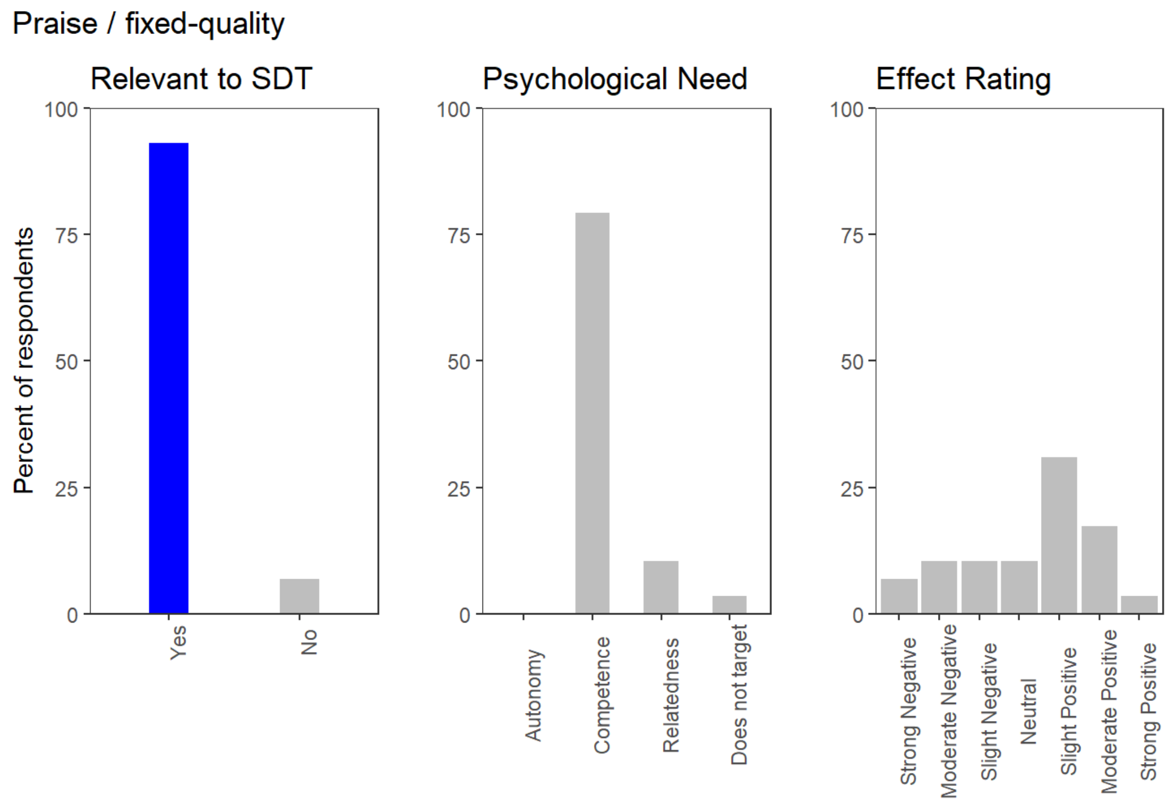
Provides praise that targets the talents or qualities of the individual

Example Behaviour:

"You are very good at maths"

Function Description:

Affirms students natural abilities



***

TMB#24

Praise / improvement or effort

Description:

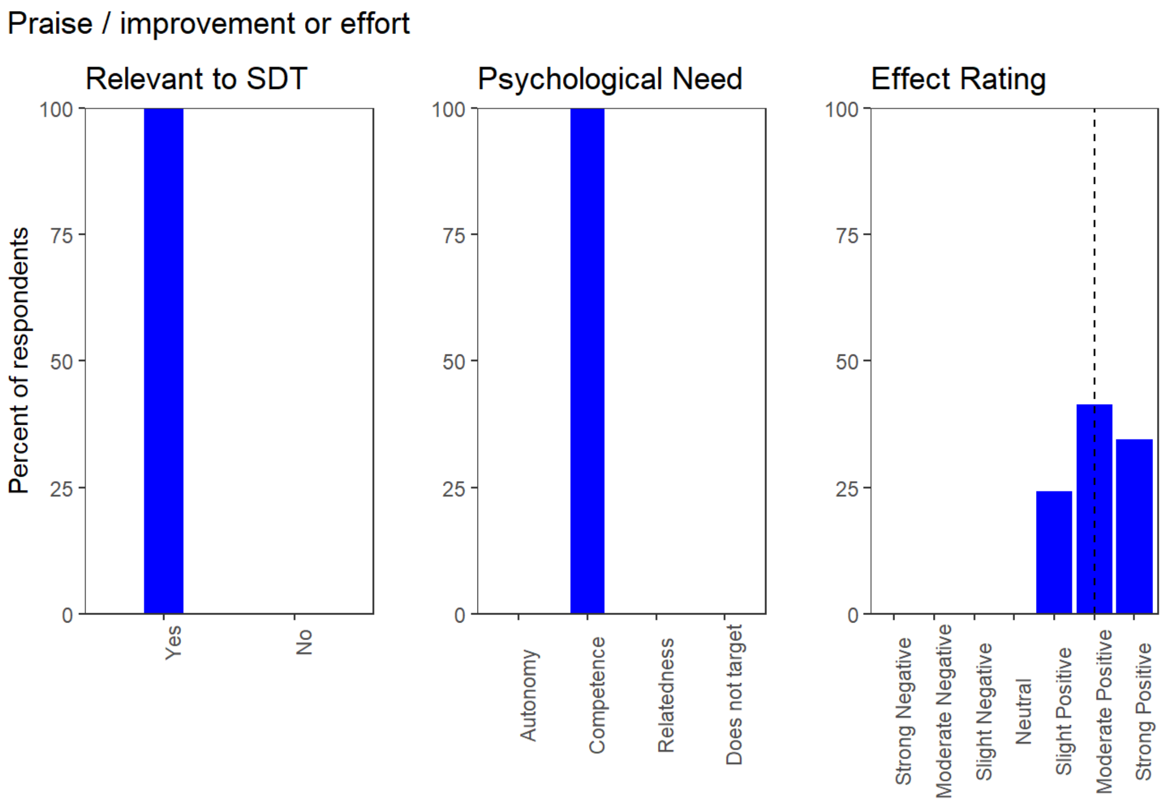
Provides praise that targets the improvement or effort from the student

Example Behaviour:

"I see some excellent hard work here, and some improvements over last weeks work, especially in these areas...."

Function Description:

Affirms students progress and improvement



***

TMB#25

Praise / public

Description:

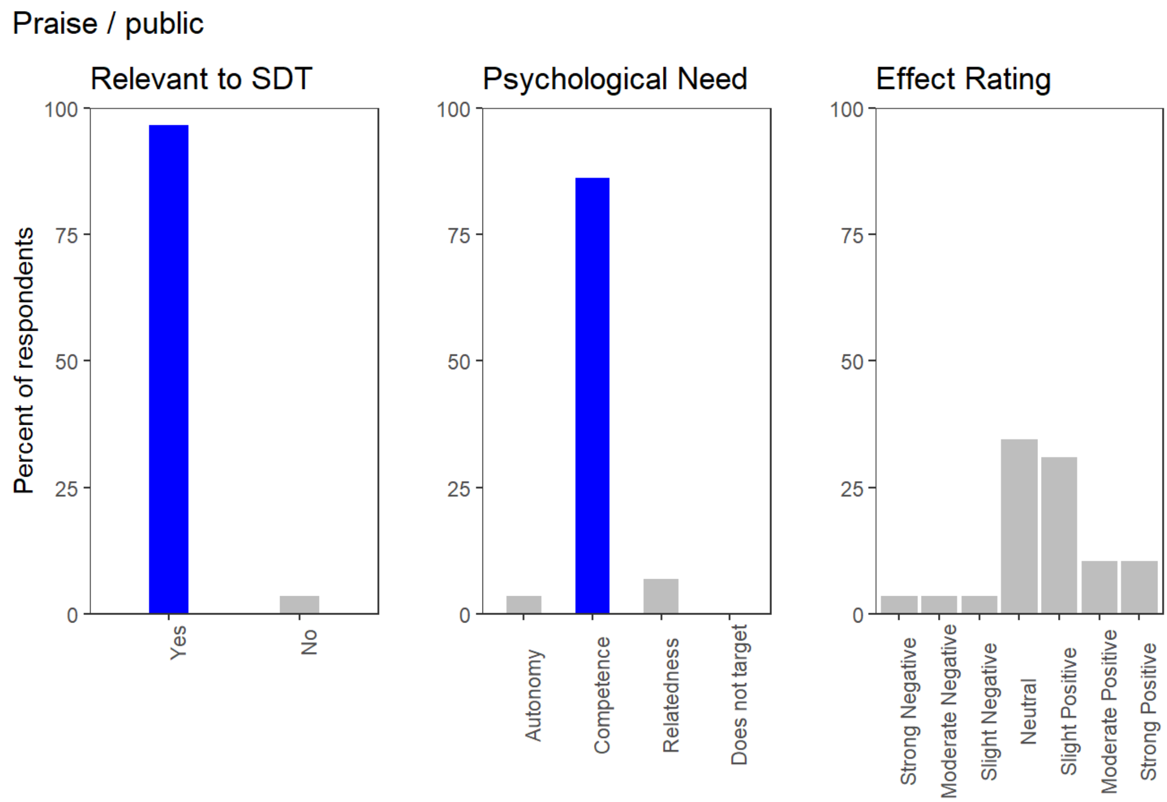
Praise (any kind of prase) a student in public

Example Behaviour:

Praise in front of the class

Function Description:

Generates pride within students receiving praise



***

TMB#26

Praise / specific

Description:

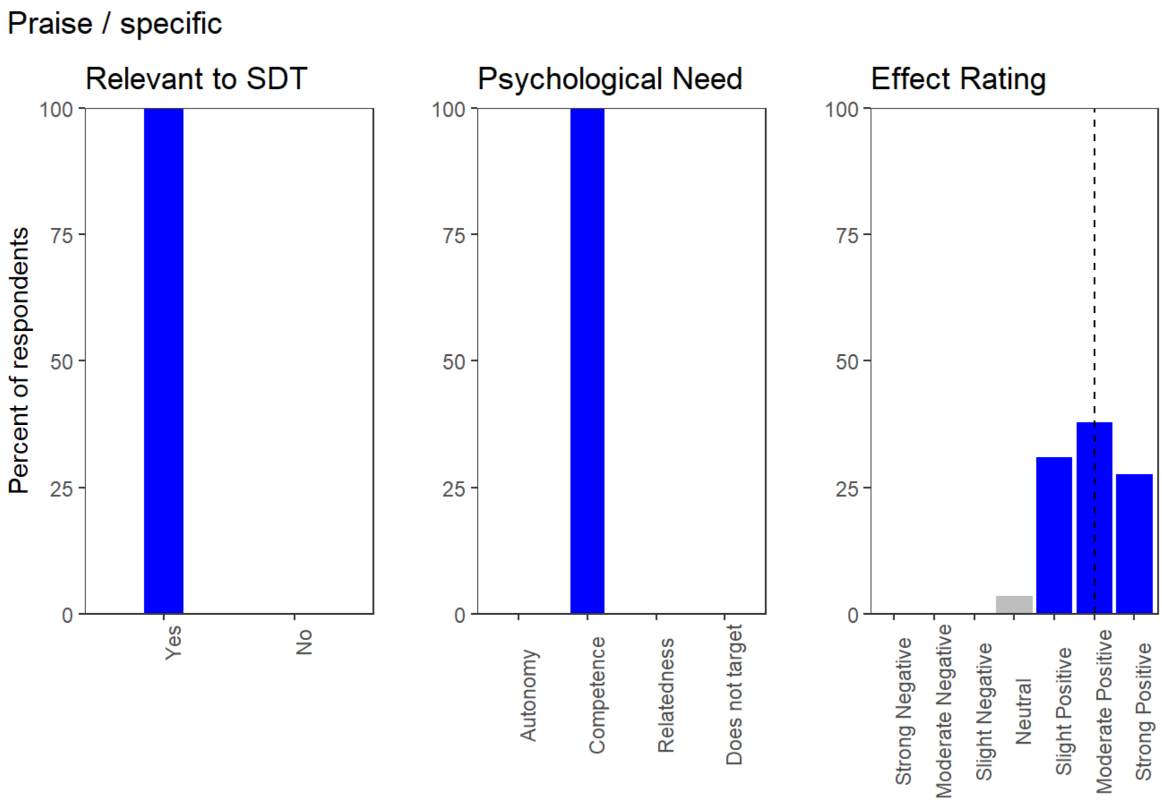
Provides praise that is specific to an action or quality of the student

Example Behaviour:

"This answer was very good because it showed the working out in clear steps"

Function Description:

Clarifies behaviours that, if repeated, lead to goal achievement



***

TMB#27

Frequent praise

Description:

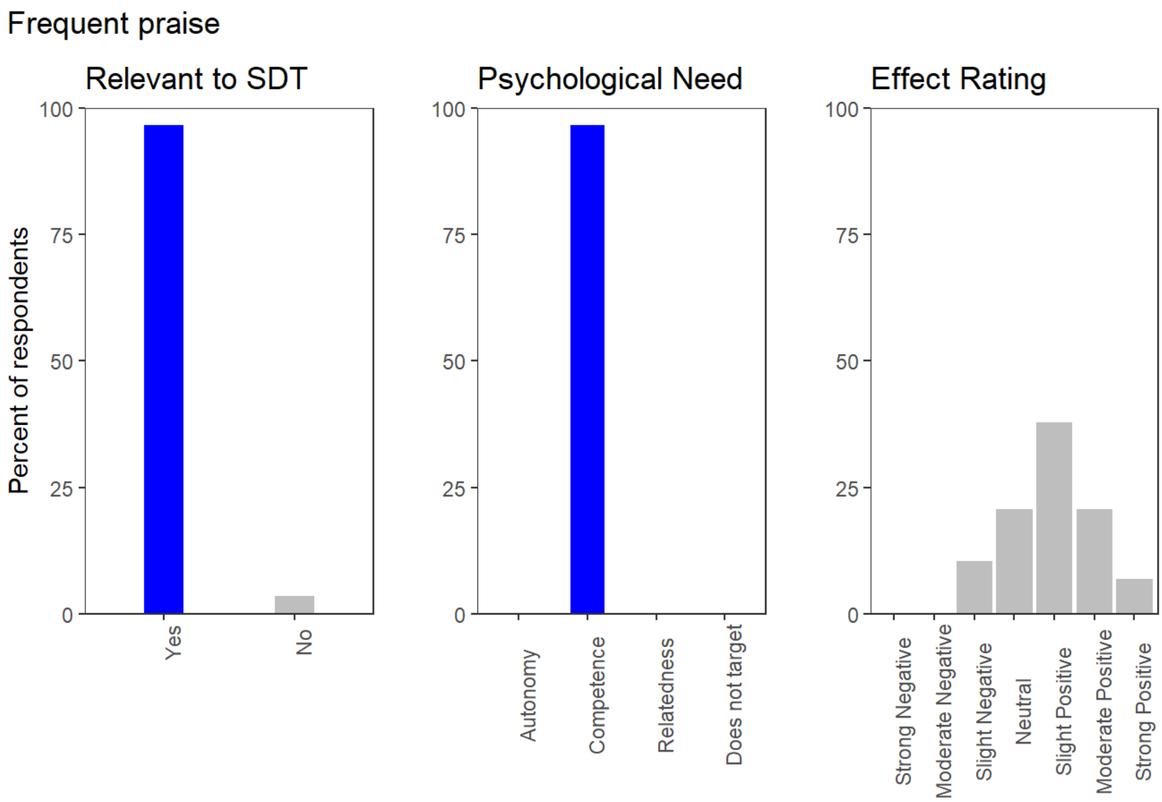
Frequently praise students for good work

Example Behaviour:

Teacher consistently patrols the class, identifying correct answers

Function Description:

Provides continual affirmation of progress and improvement.



***

TMB#28

Extra resources

Description:

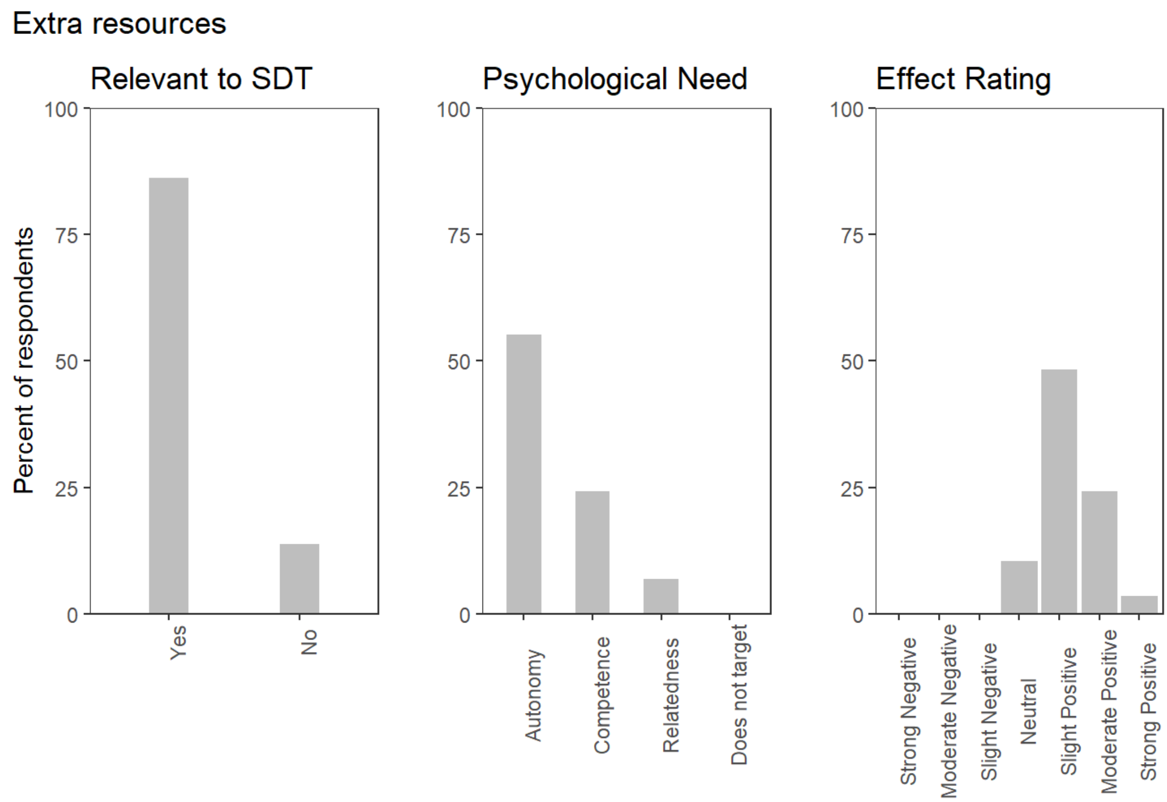
Provide options for more support or learning

Example Behaviour:

"If you want more help, remember maths club before school tomorrow"; "here are some extra problems if you want to practice at home"

Function Description:

Students can choose to get more help if they want it



***

TMB#29

Goals / self-referenced standards

Description:

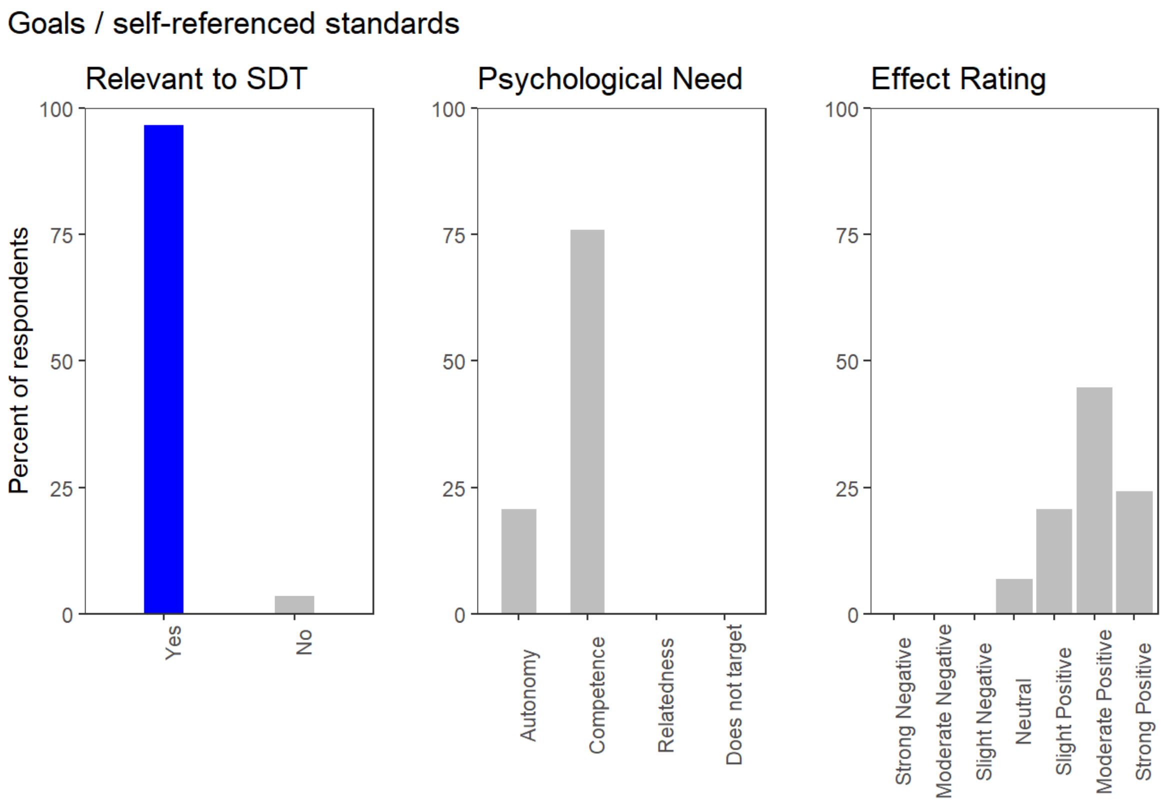
Set up activities where each student has their own goal

Example Behaviour:

"Aim to get more problems done than last time"

Function Description:

Promotes achievable goals by calibrating them to students skill



***

TMB#30

Set Heterogeneous Groups

Description:

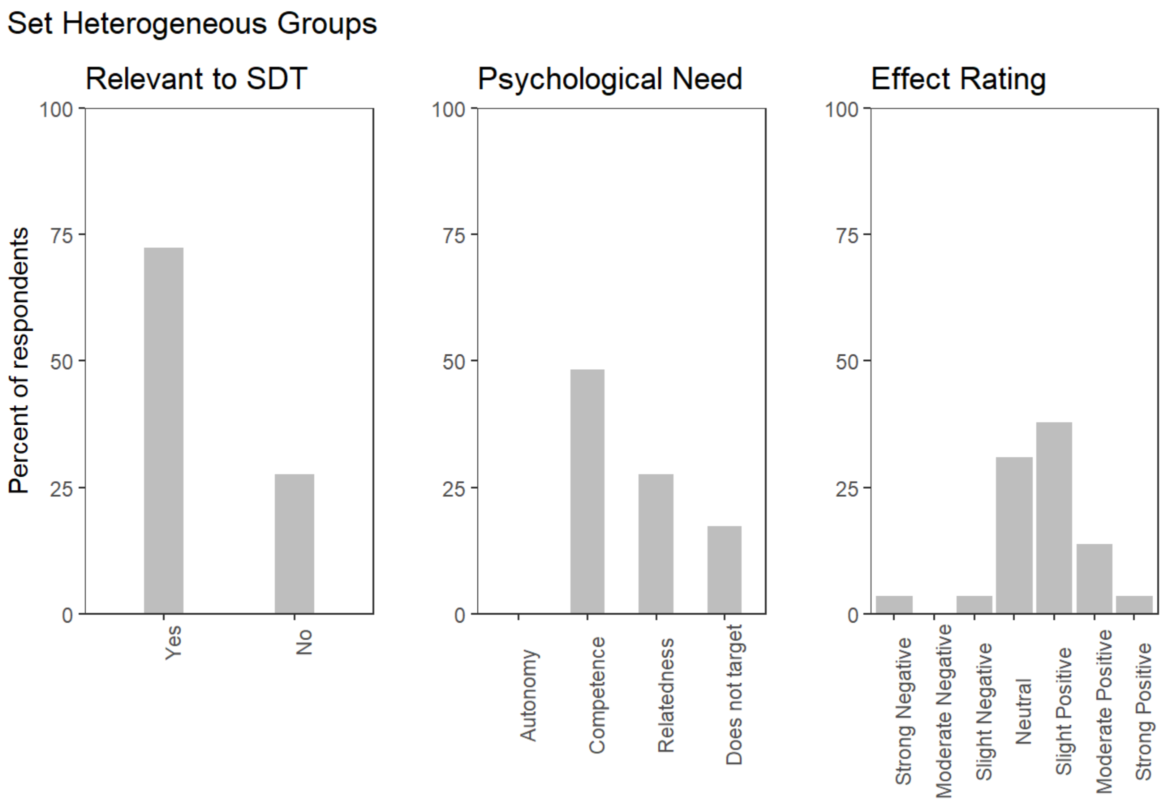
For group activities, assign students so that each group has a mix of abilities

Example Behaviour:

"Take a playing card, and find the other students with the same suit as you"

Function Description:

Removes public signalling of incompetence and ensures balanced frames of reference



***

TMB#31

Clear Instructions

Description:

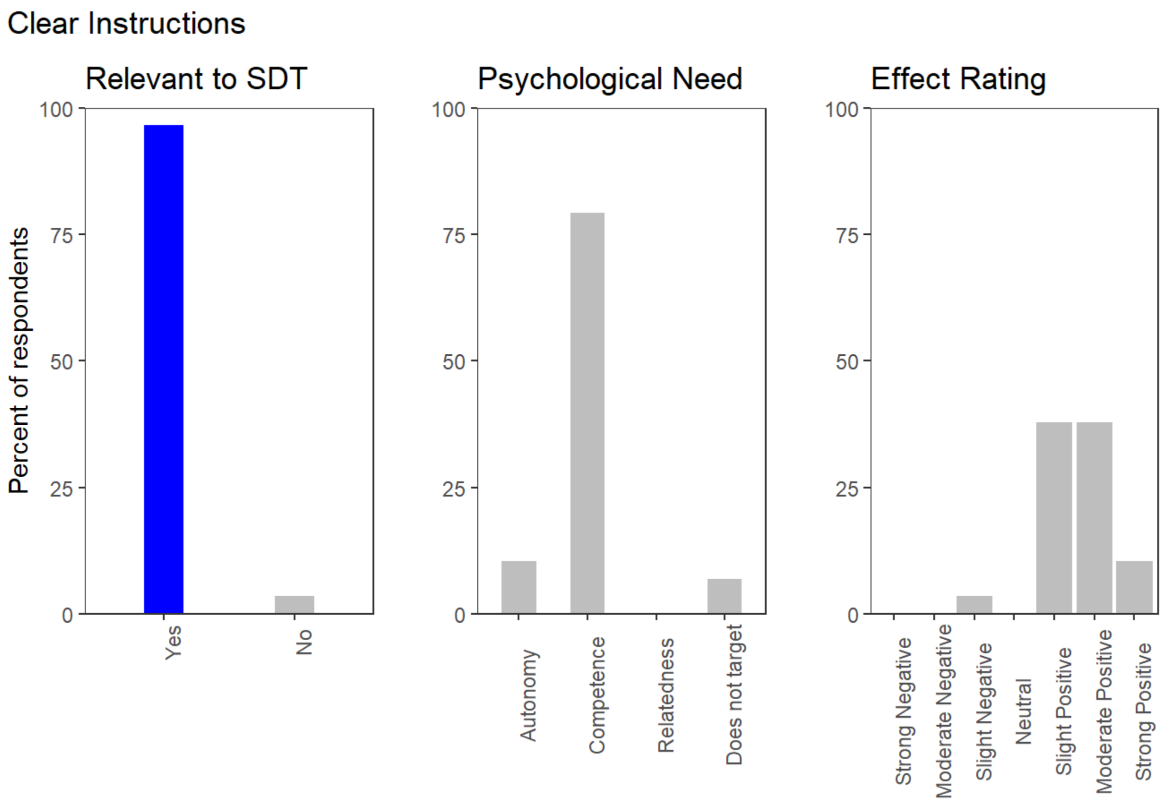
Provide clear instructions

Example Behaviour:

"Start with problems 4.1 to 4.4 then check your answers with me"

Function Description:

Students know exactly what to do



***

TMB#32

Variety

Description:

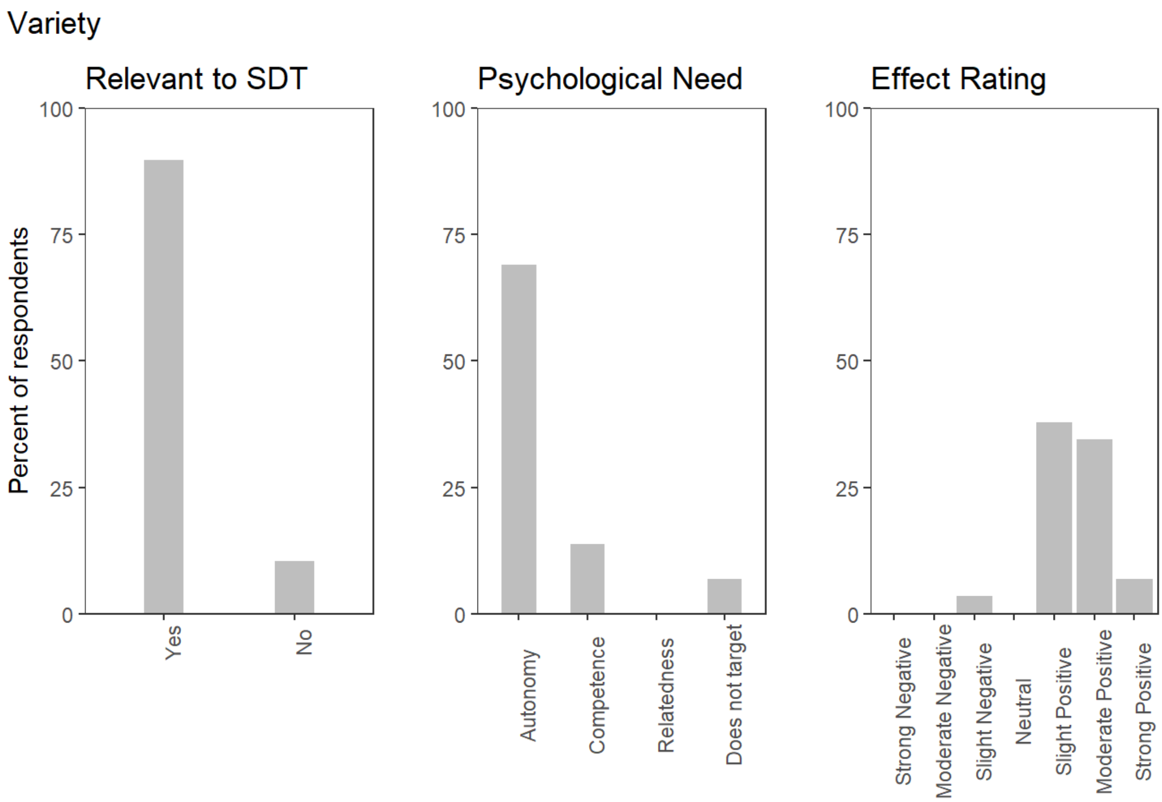
Provide a variety of activities in a way that keeps things interesting

Example Behaviour:

Teacher starts a class by giving a bit of information followed by a fun game and challenging problem

Function Description:

Reduces boredom



***

TMB#33

Transparent Structure

Description:

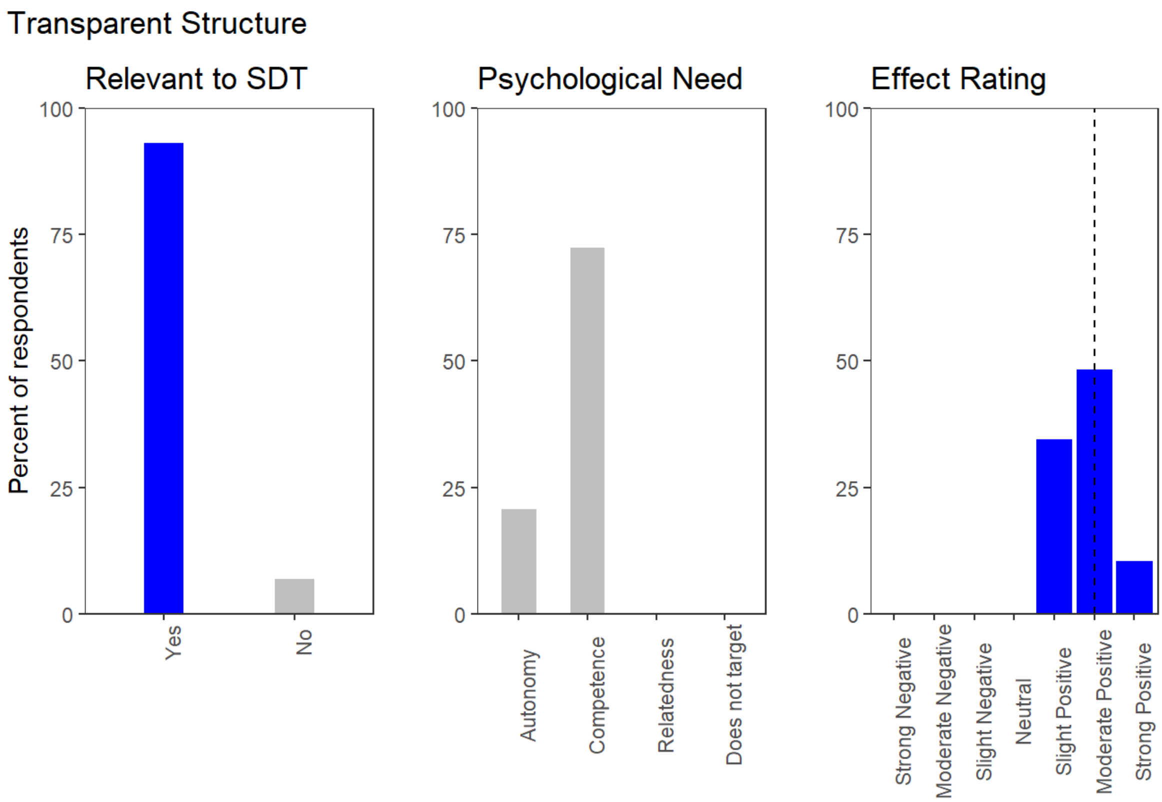
Provide an overview of what we are going to do in the lesson

Example Behaviour:

"In todays class, we are working on ratios in three ways..."

Function Description:

Students know how things are organised



***

TMB#34

Questions to check knowledge

Description:

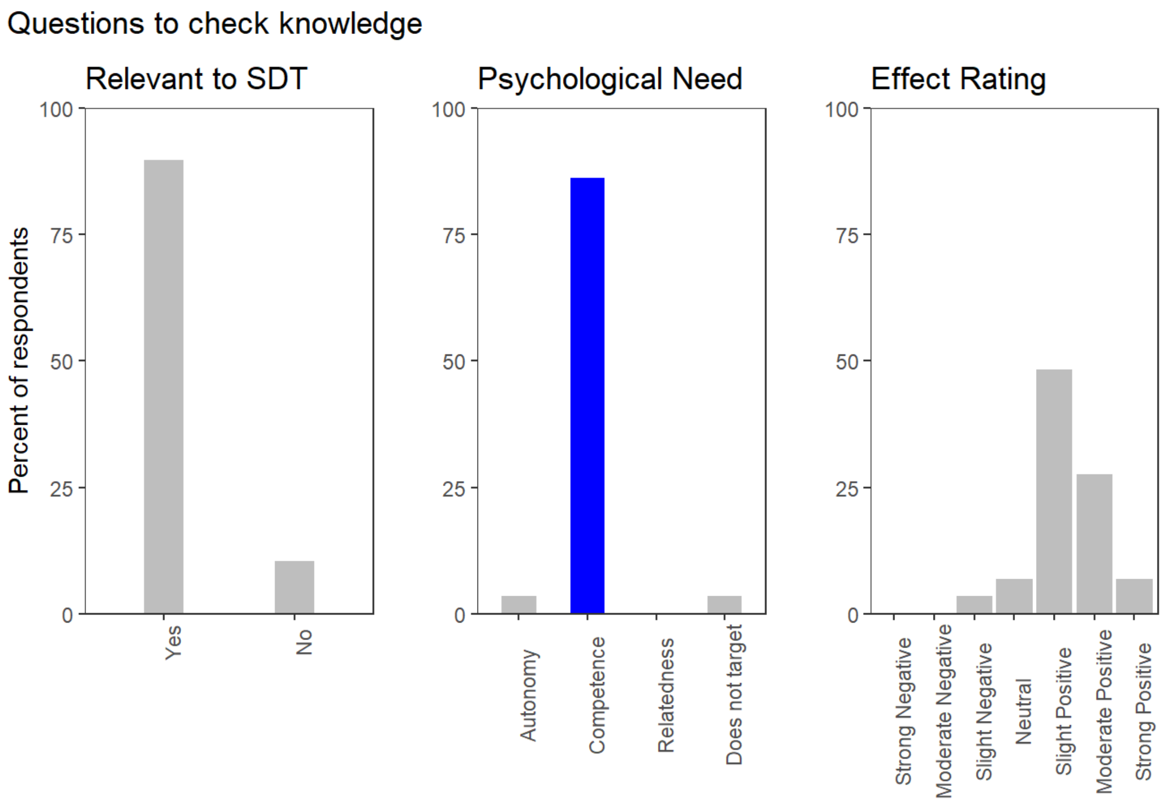
Ask the students clarification questions that check what students know

Example Behaviour:

"What is the B in BODMAS?"

Function Description:

Fosters common understanding of goal-directed behaviours



***

TMB#35

Questions to expand understanding

Description:

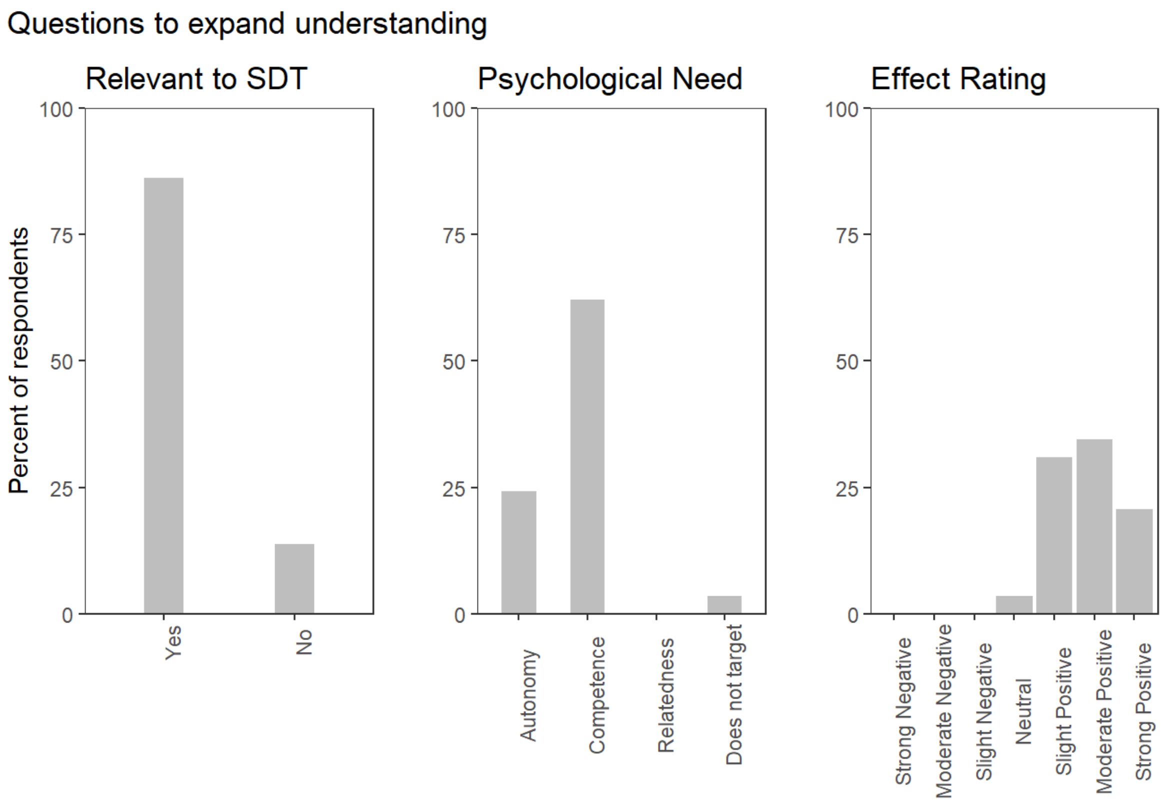
Questioning to expand understanding or thinking

Example Behaviour:

PE - "What other sports do we use these skills?", Maths - "When might we use division in our daily lives?"

Function Description:

Fosters deeper understanding of how knowledge fits together



***

TMB#36

Responding to Queries

Description:

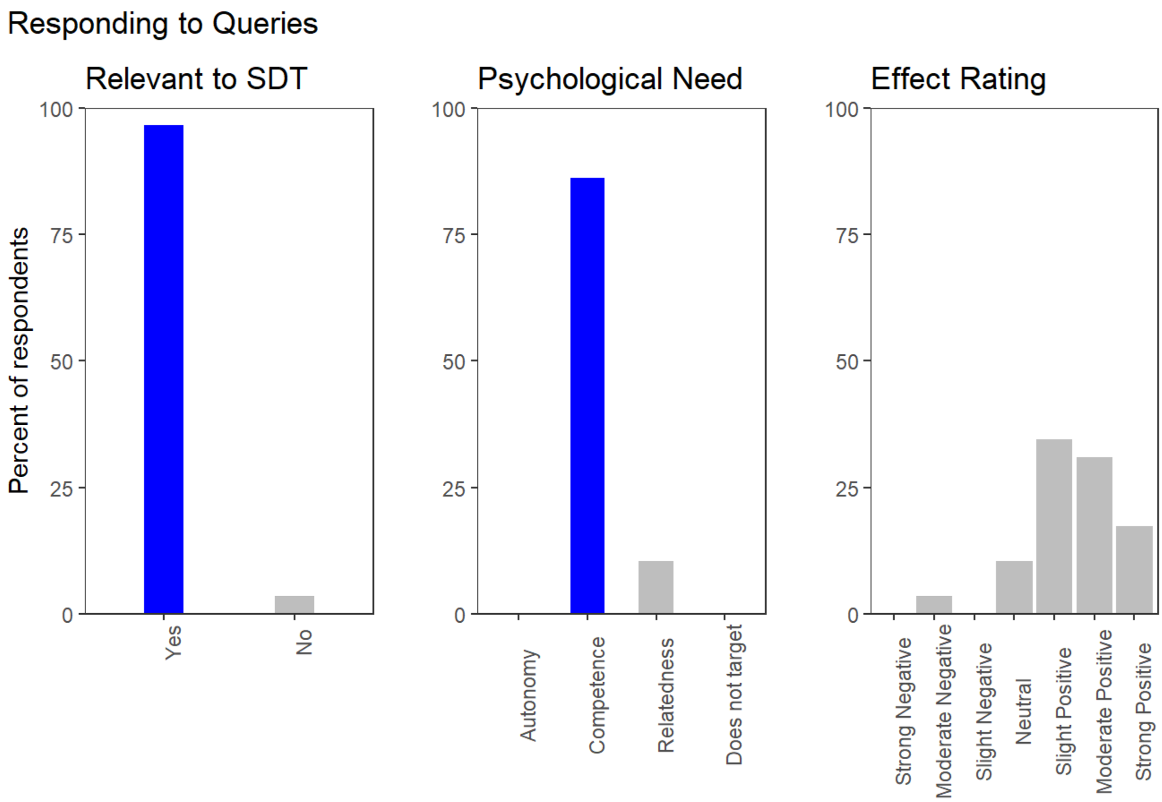
Answer student questions fully and carefully

Example Behaviour:

"No, that is the formula for Sin not Cos"

Function Description:

Clarifies path toward goal achievement.



***

TMB#37

Communicating perspective-taking statements

Description:

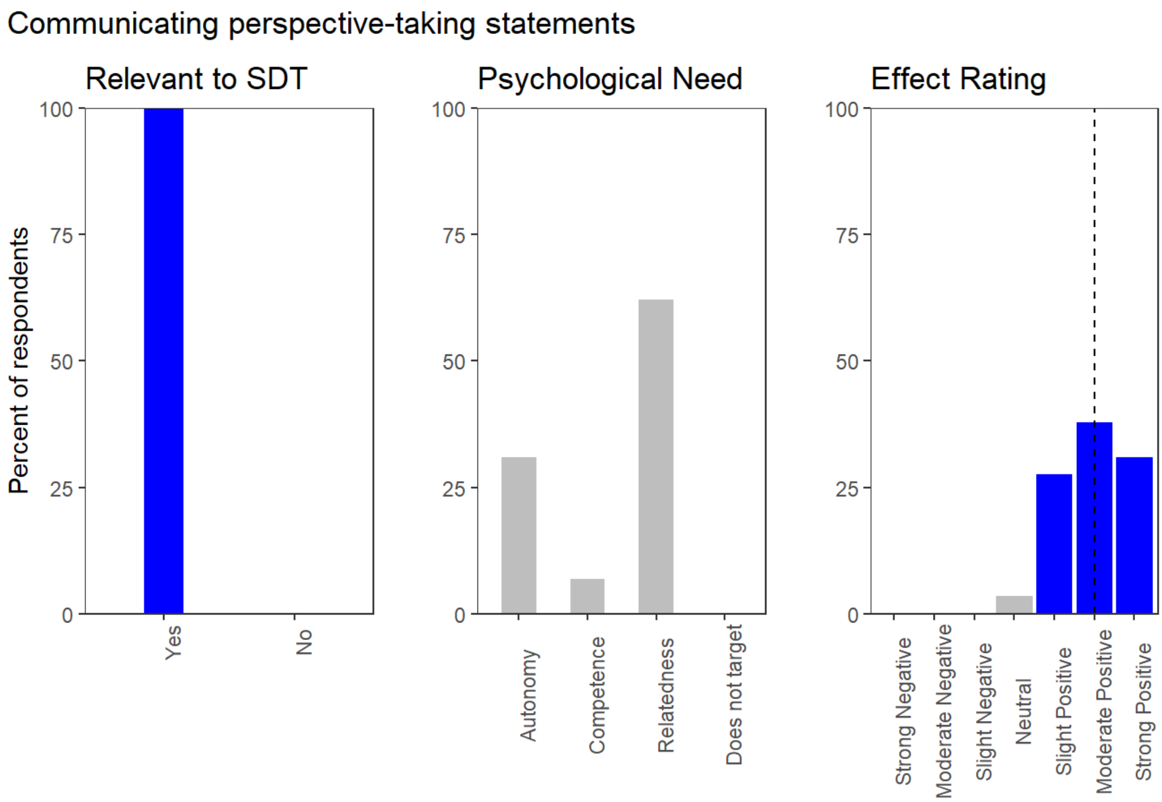
Show that you have taken a students perspective

Example Behaviour:

“Yes, this one is difficult”; “I know it is a sort of difficult one.”

Function Description:

Communicates that teacher understands the students frame of reference



***

TMB#38

Show Interest in Outside School activities

Description:

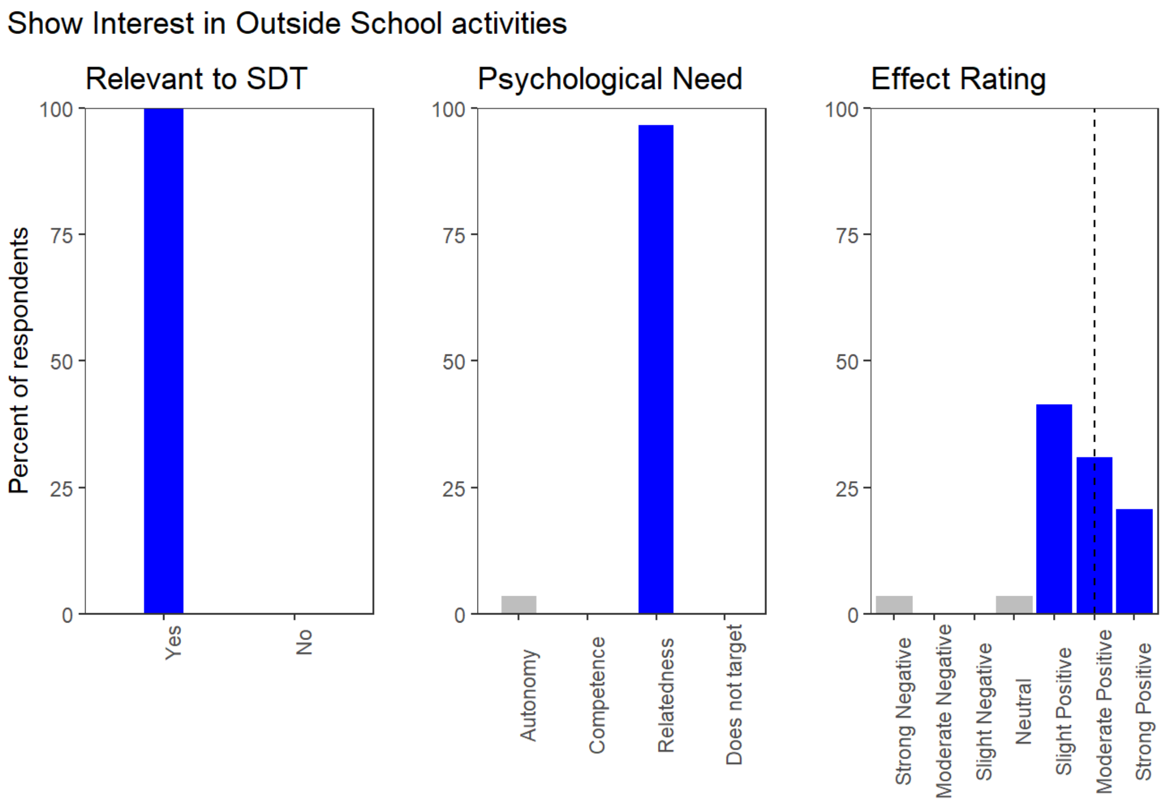
Express interest in students activities outside of school

Example Behaviour:

"What sports did you do on the weekend?"

Function Description:

Enables students to feel important and connected



***

TMB#39

Acknowledge student negative feelings

Description:

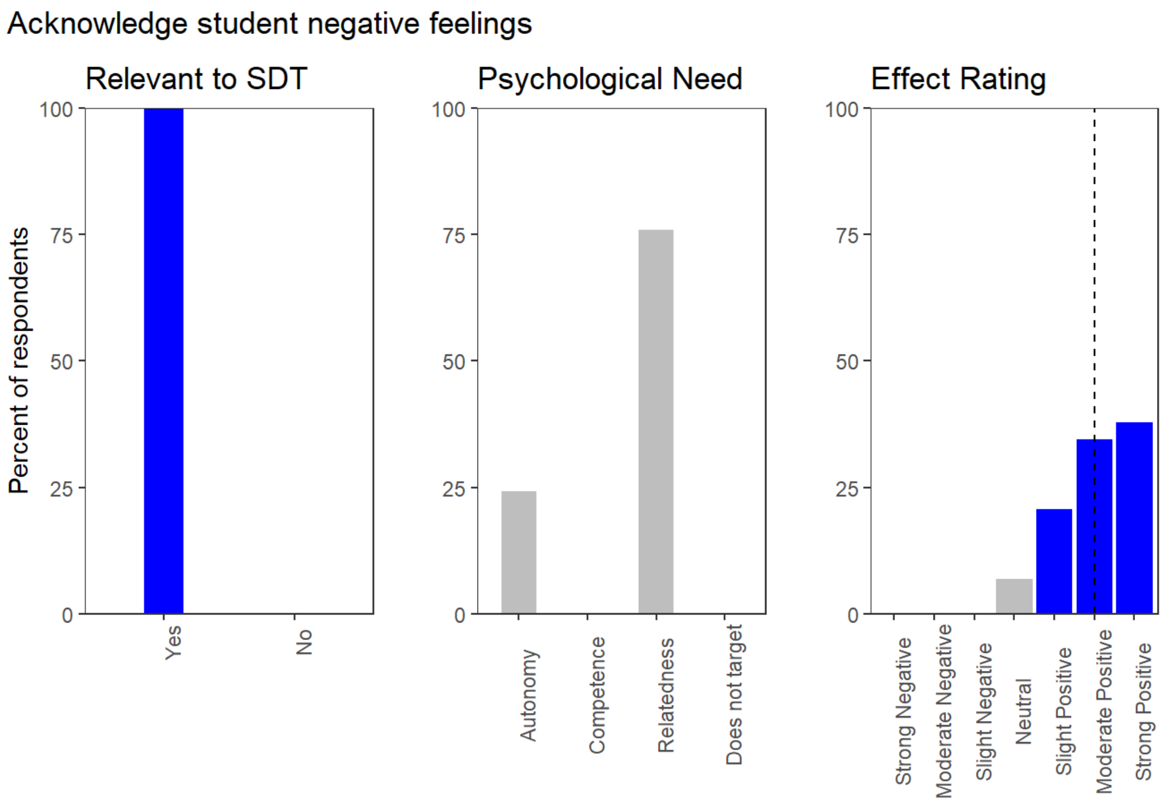
Acknowledge students negative feelings

Example Behaviour:

"I noticed you are looking frustrated."

Function Description:

Validates emotions as understandable, normal, and expected



***

TMB#40

Rely on invitational language

Description:

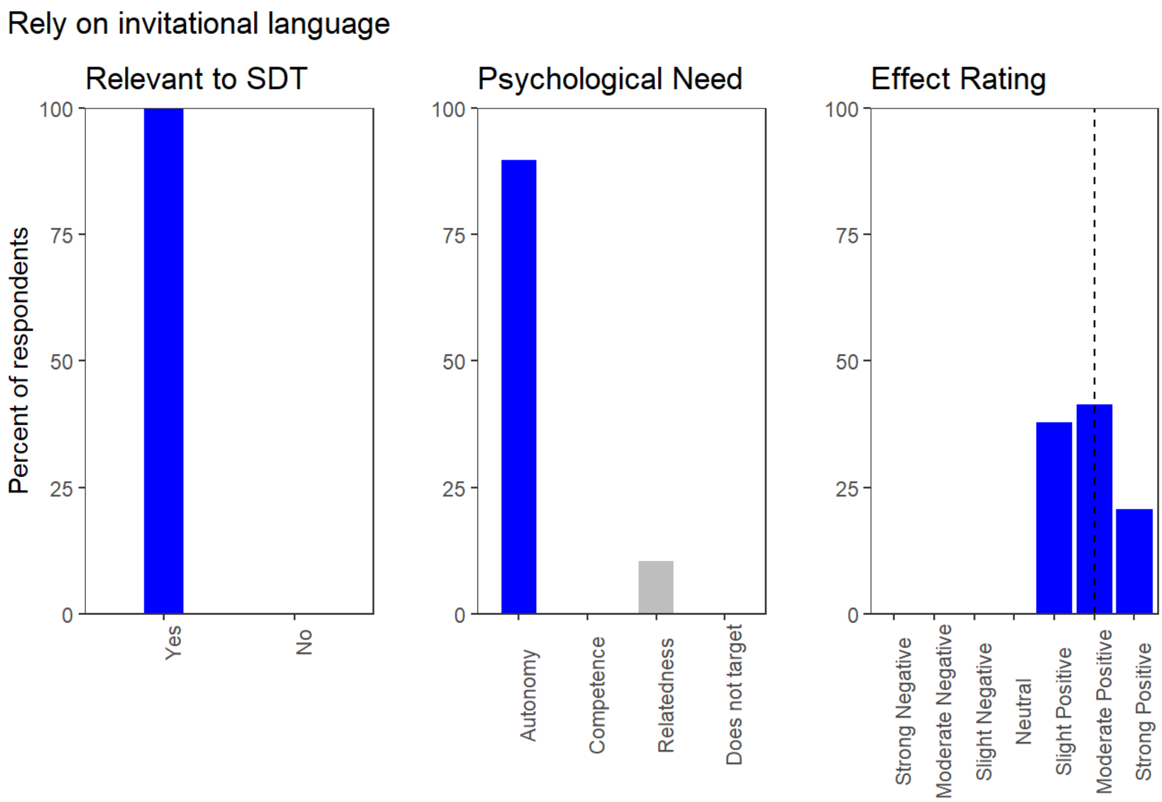
Instead of telling students what they must, have to, or should do, invite students to self-initiate into learning activities

Example Behaviour:

"You may want to try this... " and "This behavior has worked for students in the past who have had this same problem, what do you think?"

Function Description:

Reduces perceived external pressure to complete the task for imposed reasons.



***

TMB#41

Allow student own-paced progress

Description:

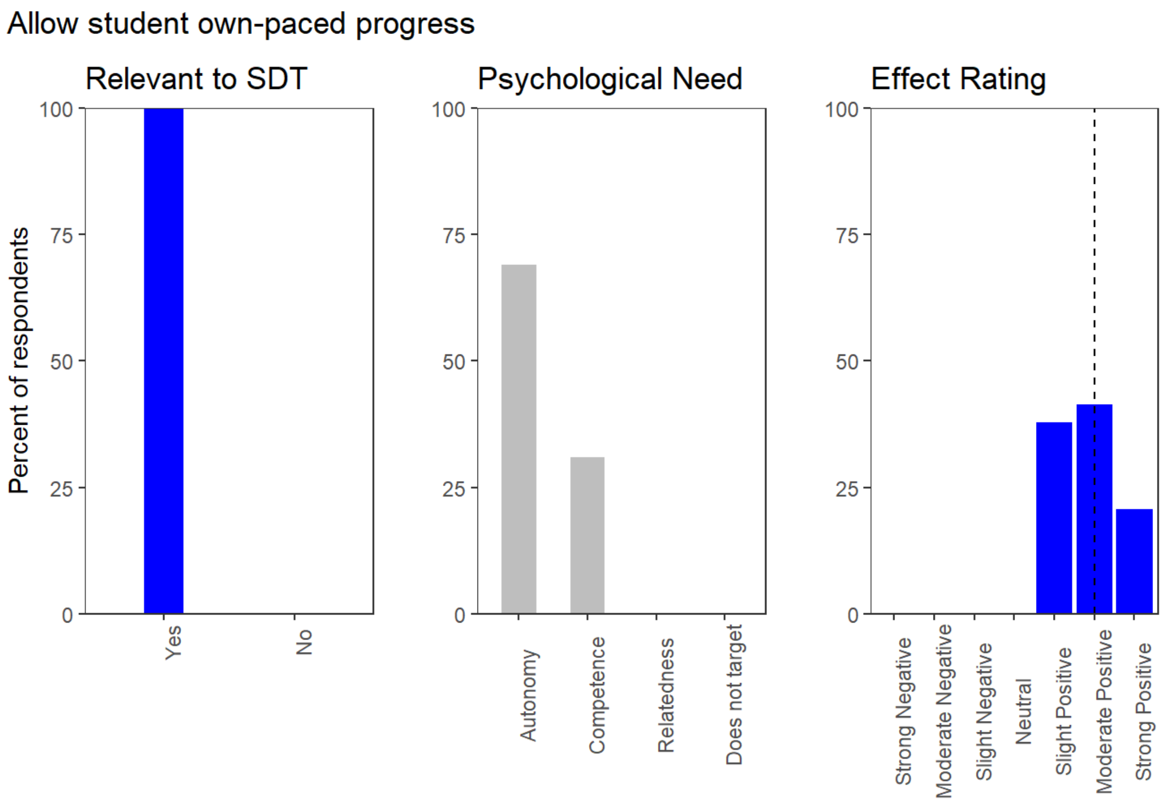
Allow the student to work independently and to solve a problem in his or her own pace

Example Behaviour:

"Solve the puzzle at your own pace"

Function Description:

Lets students manage their own cognitive load so they do not get frustrated or overwhelmed



***

TMB#42

Teaching in students' preferred ways

Description:

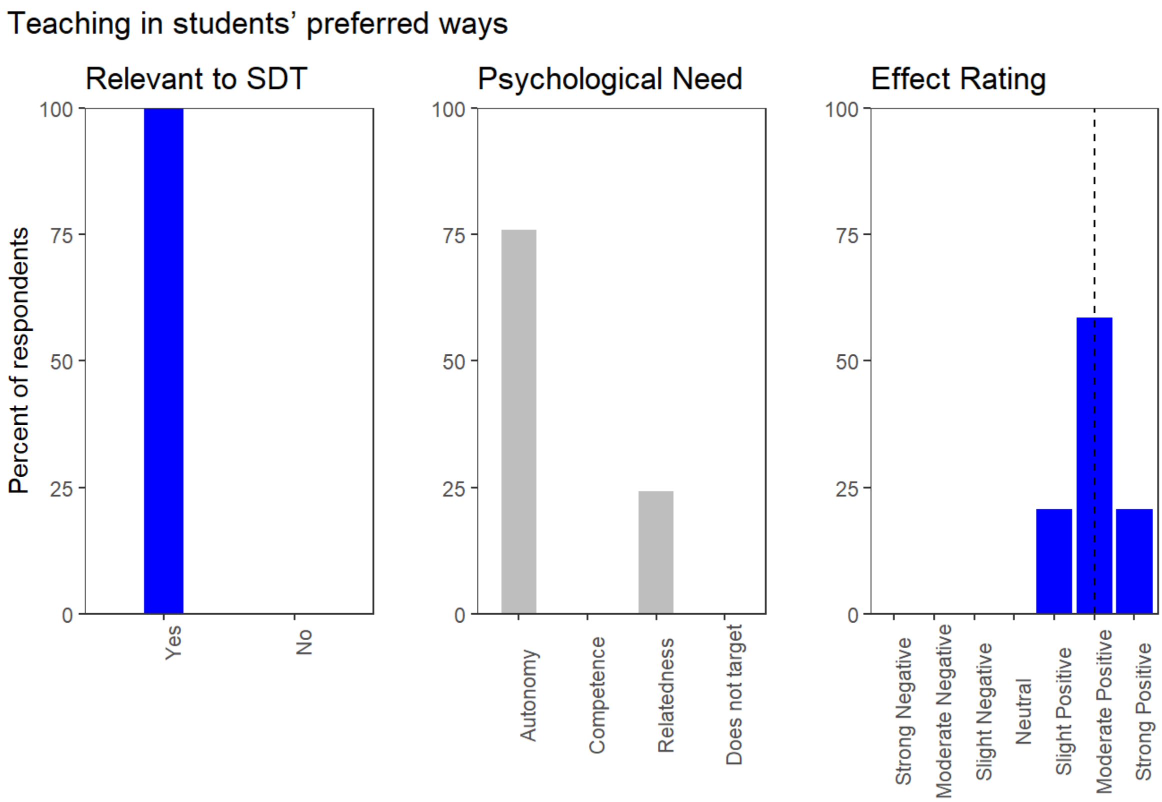
Use knowledge gleaned about the student values and preferences to design class activities customised to them.

Example Behaviour:

"I know you love comics so I based todays lesson on ..."

Function Description:

Aligns lesson activities to students intrinsic reasons for learning rather than imposing extrinsic reasons



***

TMB#43

Curiosity

Description:

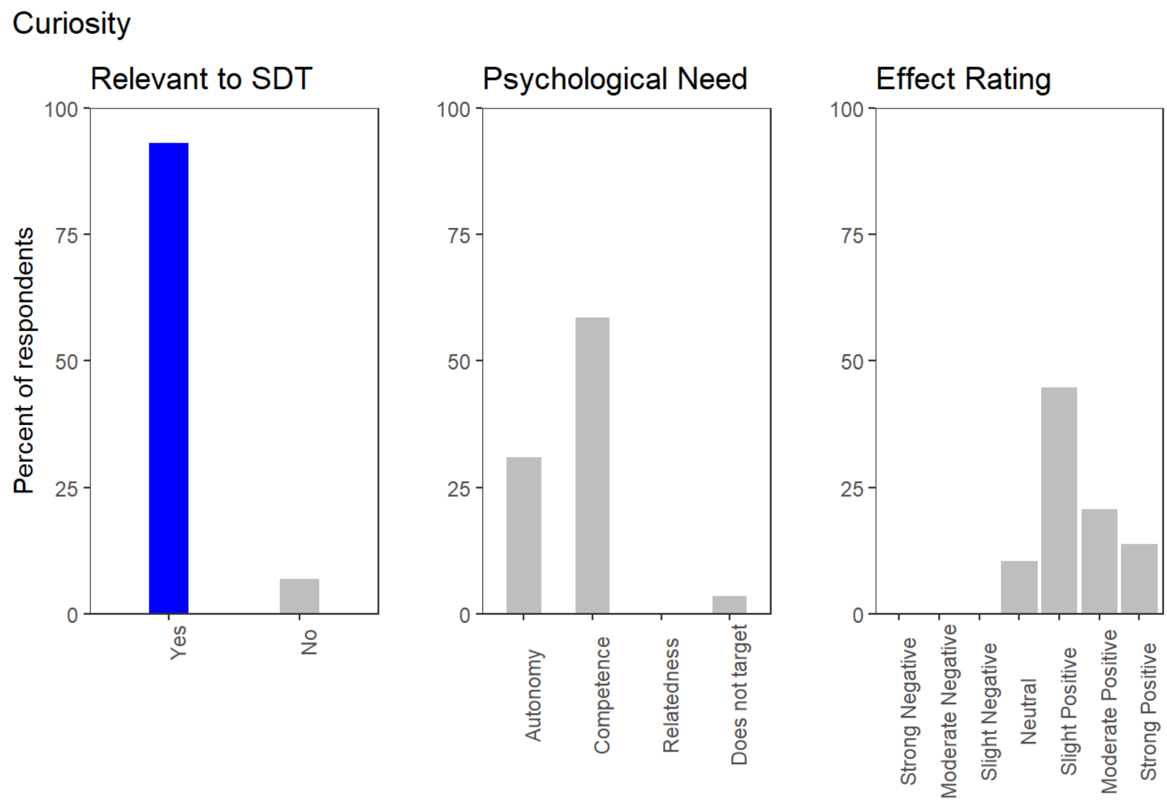
Ask a curiosity-inducing question

Example Behaviour:

"How long does it take the Earth to go around the sun?"

Function Description:

Supports students competence through facilitating their exploratory behaviour



***

TMB#44

Display explicit strong guidance

Description:

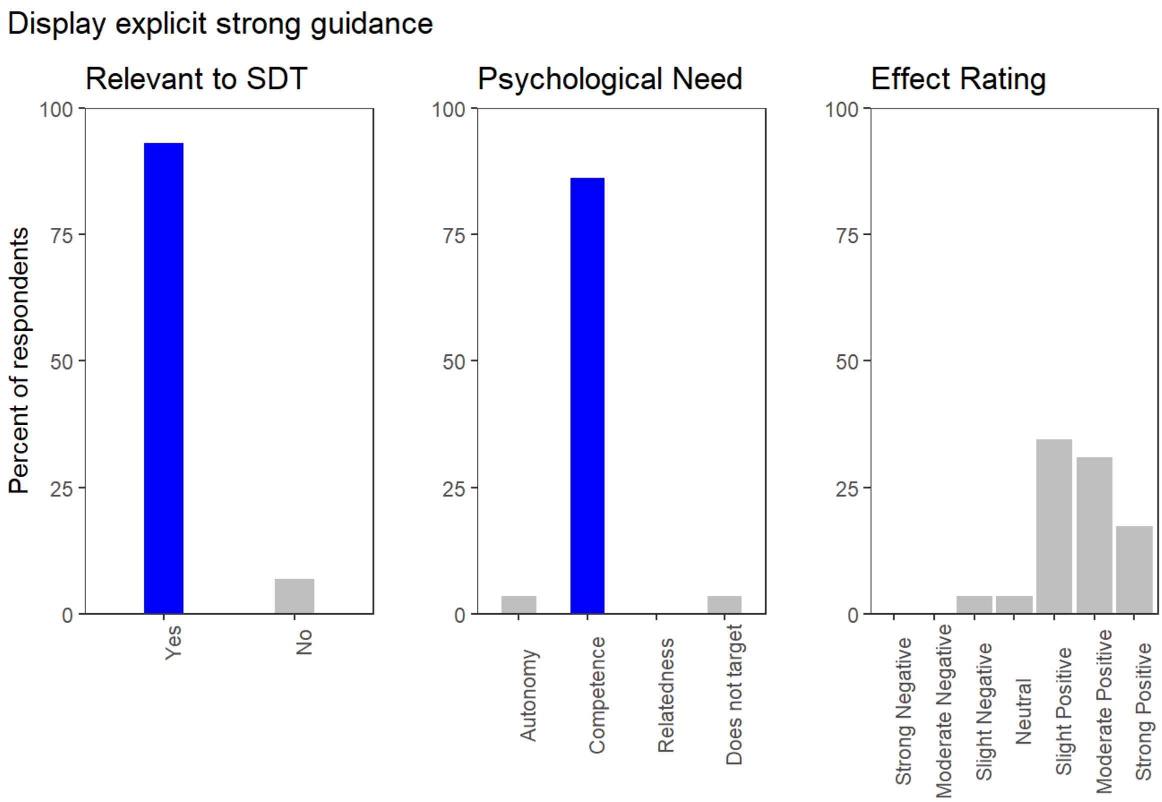
Provide clear guidance, clear goal, and clear action plans

Example Behaviour:

"Today we want to understand how volcanoes work. To do this, we are going to..."

Function Description:

Enables students to understand success criteria



***

TMB#45

See/understand from students point of view

Description:

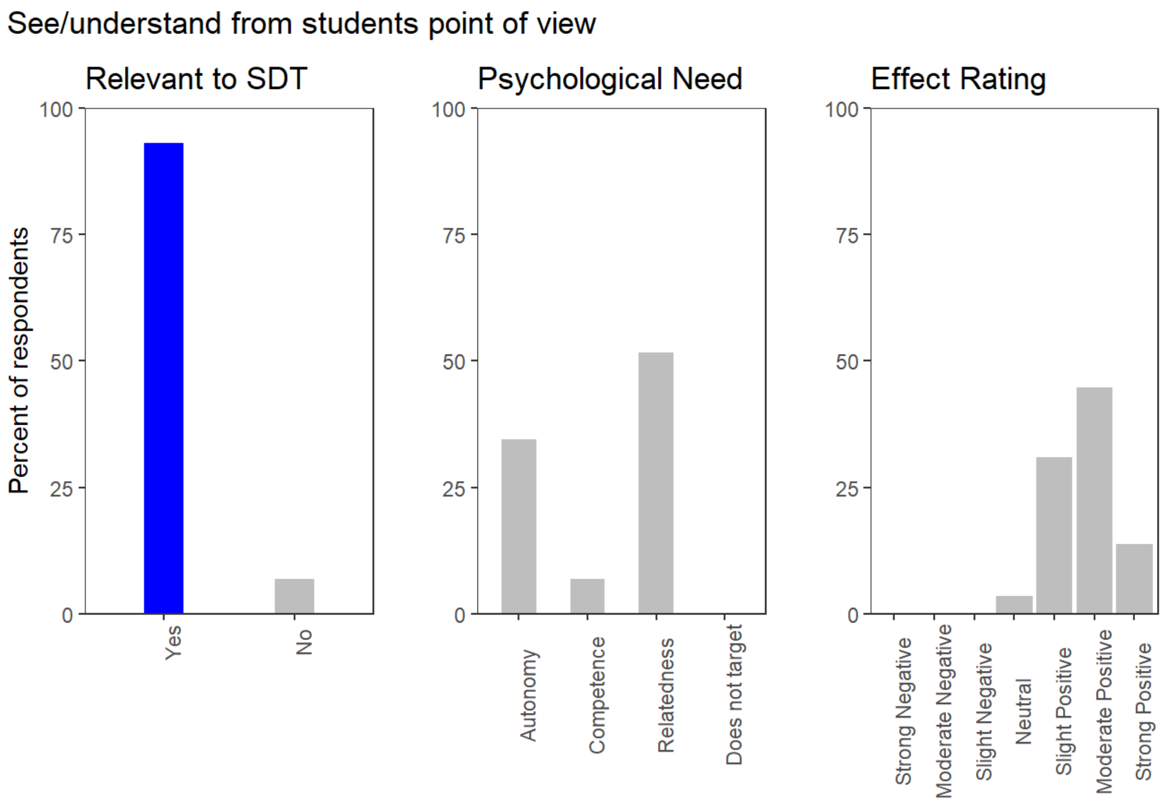
Try to understand how students see things before suggesting a new way to do things.

Example Behaviour:

"Many students think it takes a day for the Earth to go around the sun."

Function Description:

Helps the student feel listened-to and understood.



***

TMB#46

Use pupils as positive role models

Description:

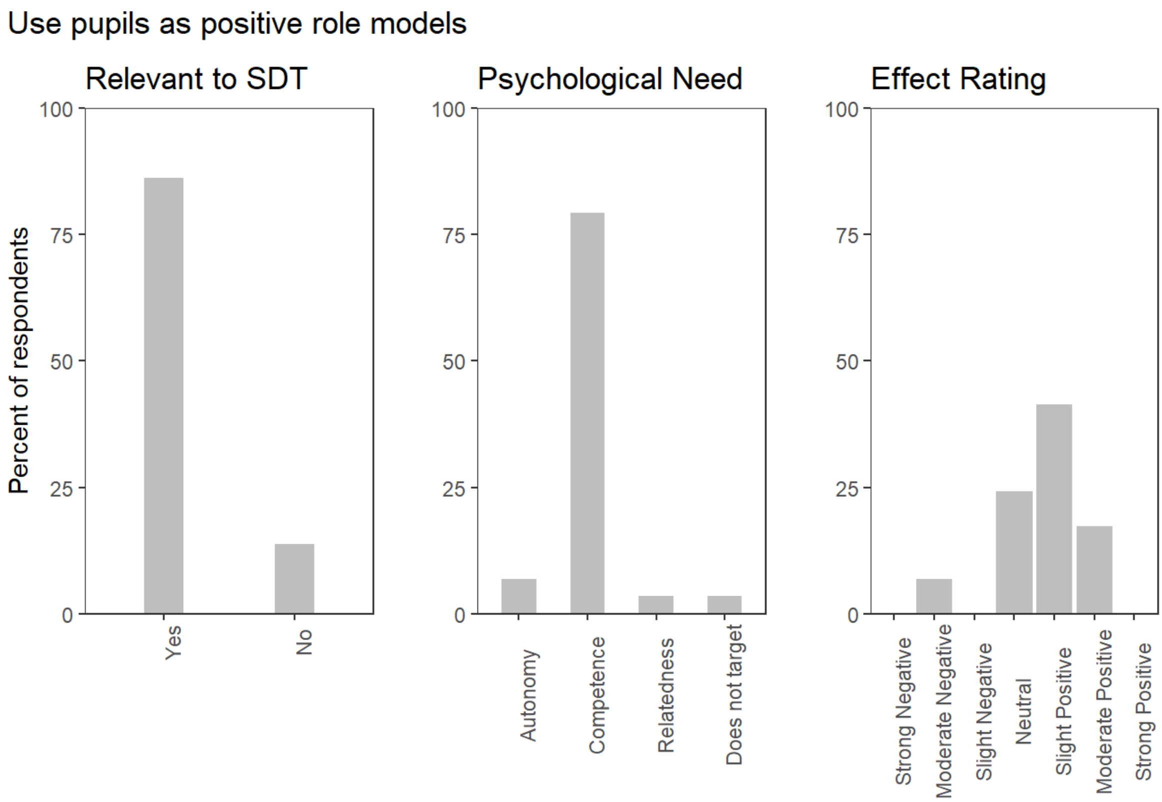
Use pupils as positive role models

Example Behaviour:

"John, you did the technique well. Can you do it again so your friends can see it?"

Function Description:

Increase self-belief through vicarious experiences of success



***

TMB#47

Promote self-assessment

Description:

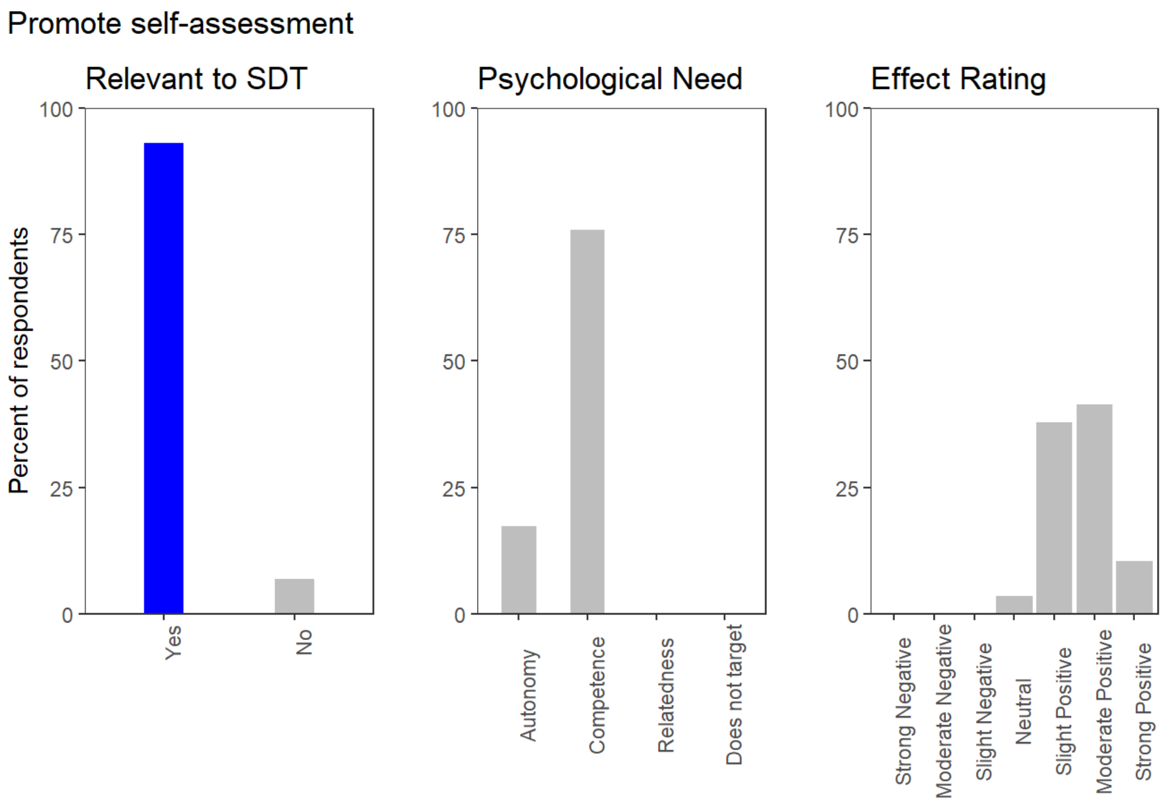
Facilitate monitoring of progress, skill level, or performance

Example Behaviour:

"How do you feel about your performance in the last three weeks?"

Function Description:

Provides opportunities for accurate self-reflection on competence



***

TMB#48

Prefer open-ended questions over closed questions

Description:

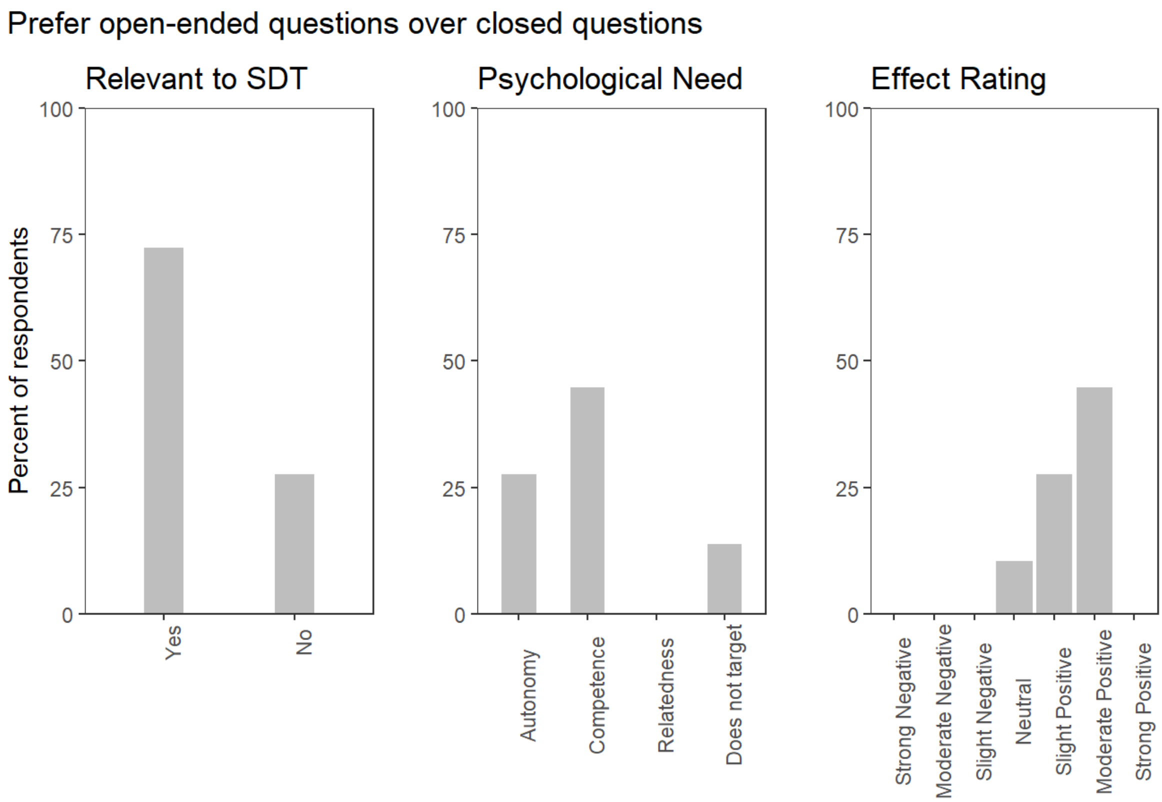
Ask questions that require many words to answer

Example Behaviour:

Ask questions starting with "why", "how", or "what" rather than "do", "is", or "are"

Function Description:

Facilitates student self-expression and deeper thinking



***

TMB#49

Undifferentiated challenge

Description:

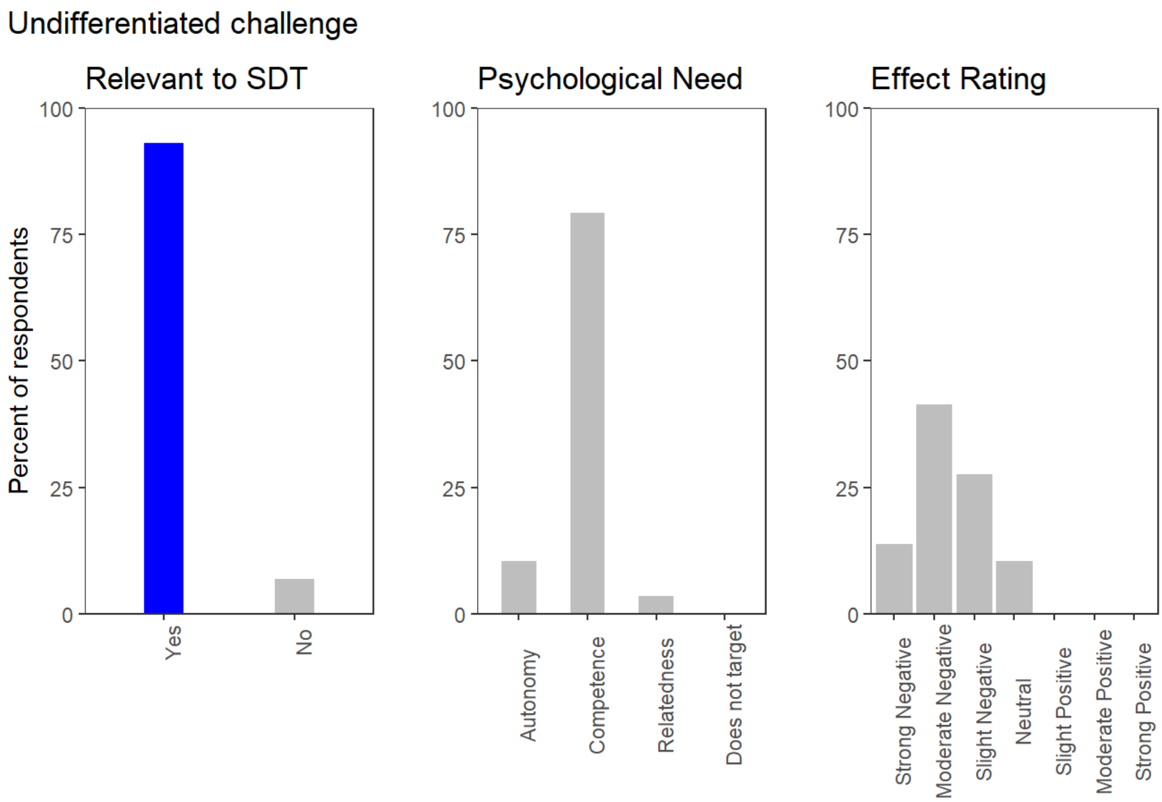
The same task is set for all students regardless of their level of ability.

Example Behaviour:

"try to do a lay up by using the backboard"

Function Description:

Given natural variation in abilities, many students may be bored and others overwhelmed.



***

TMB#50

Ignoring students

Description:

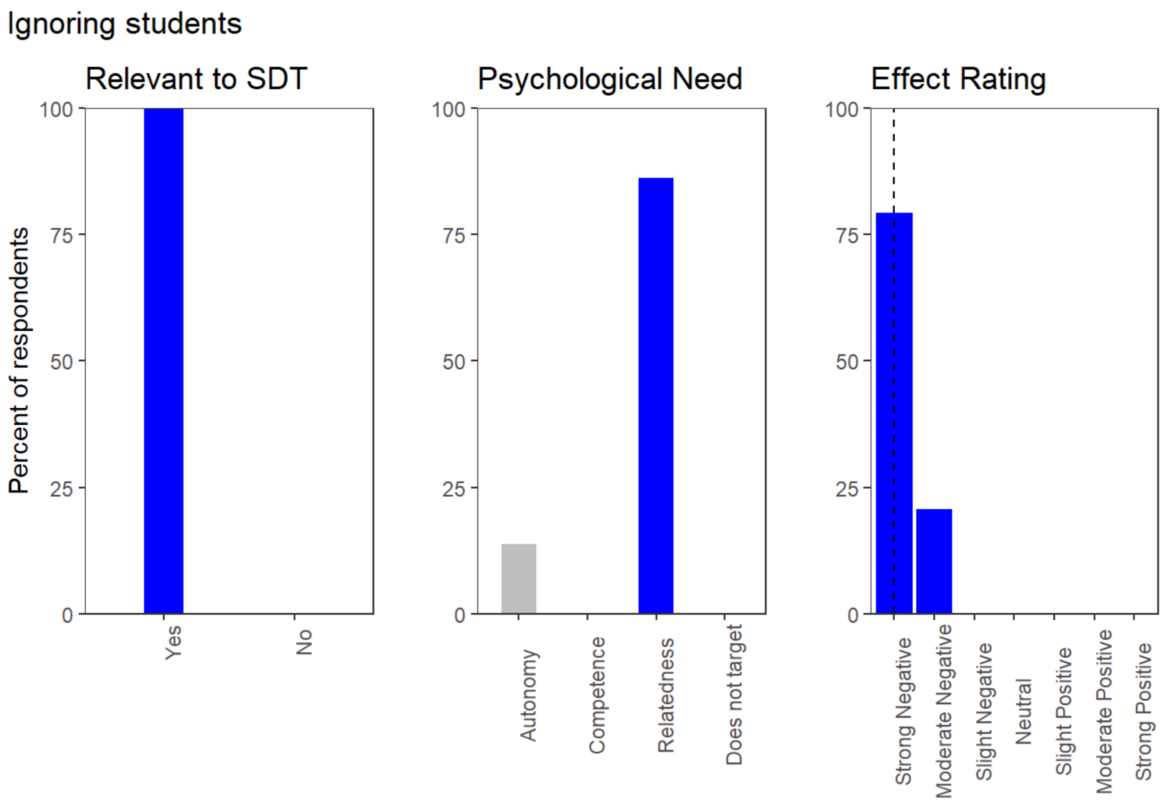
During times where attending to students would be appropriate (e.g., emotional distress, misbehaviour, active learning) the teacher maintains distance or does not direct attention to the student.

Example Behaviour:

The teacher ignores an upset student

Function Description:

Makes students feel they are not valued or cared for and that their efforts are not noticed.



***

TMB#51

Use of pressuring language

Description:

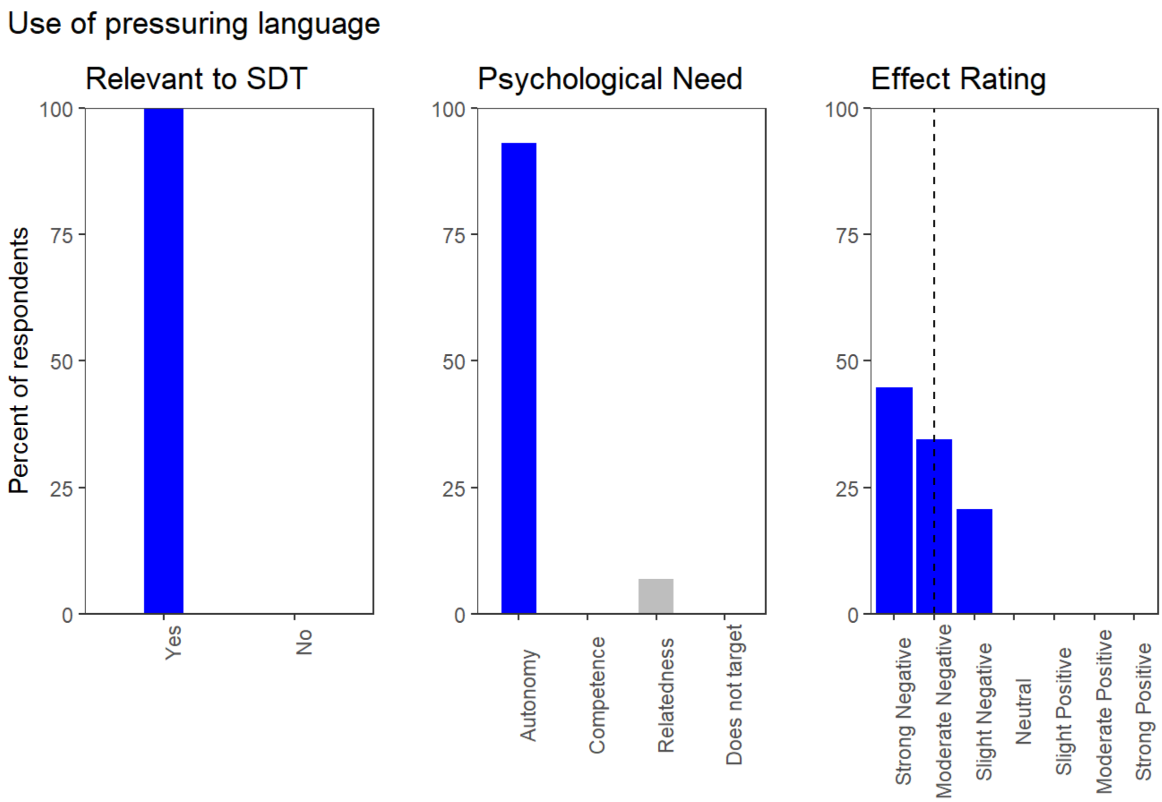
Using pressuring or controlling language when explaining tasks, providing feedback, etcera.

Example Behaviour:

"You should ...", "You have-to ...", "You must ..."

Function Description:

Increases perceived external pressure to complete the task for imposed reasons.



***

TMB#52

Uttering directives / commands

Description:

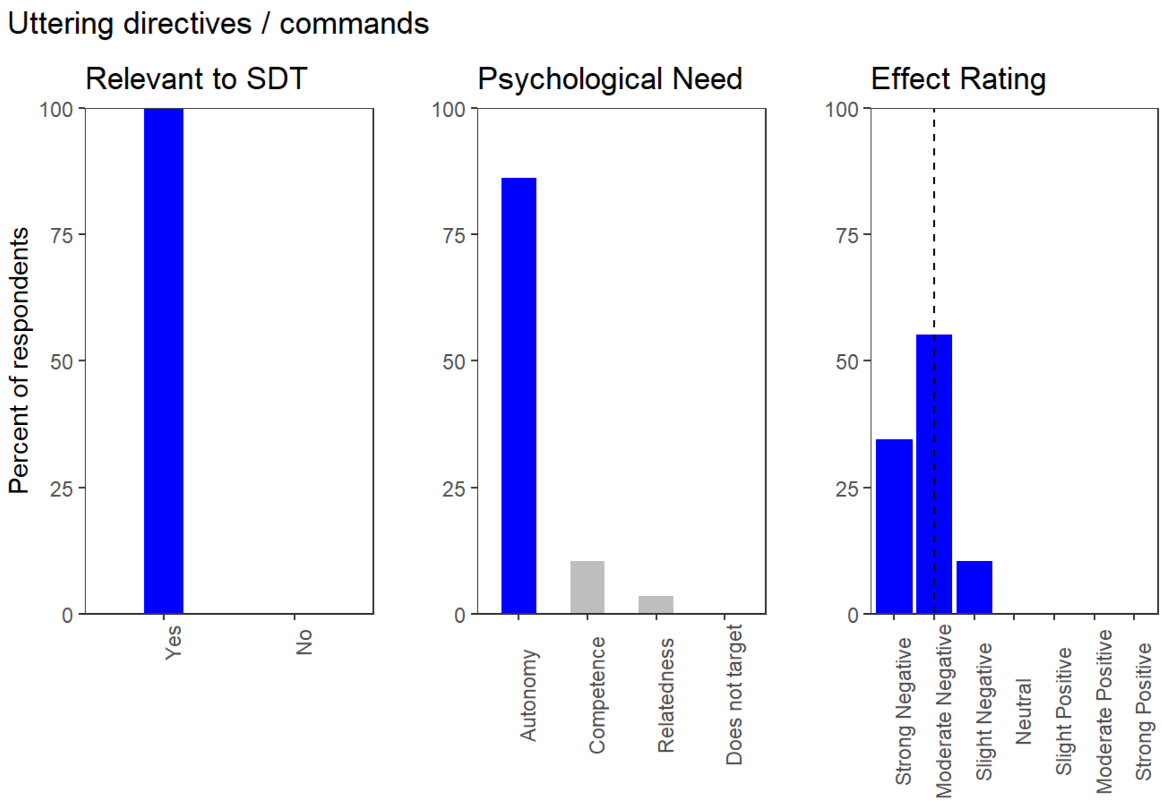
Command students to do things without providing rationales.

Example Behaviour:

“Do it like this,” “Move over here,” or “Put that away"

Function Description:

Imposes expectations on students without aligning it to any reasons.



***

TMB#53

Ask controlling questions

Description:

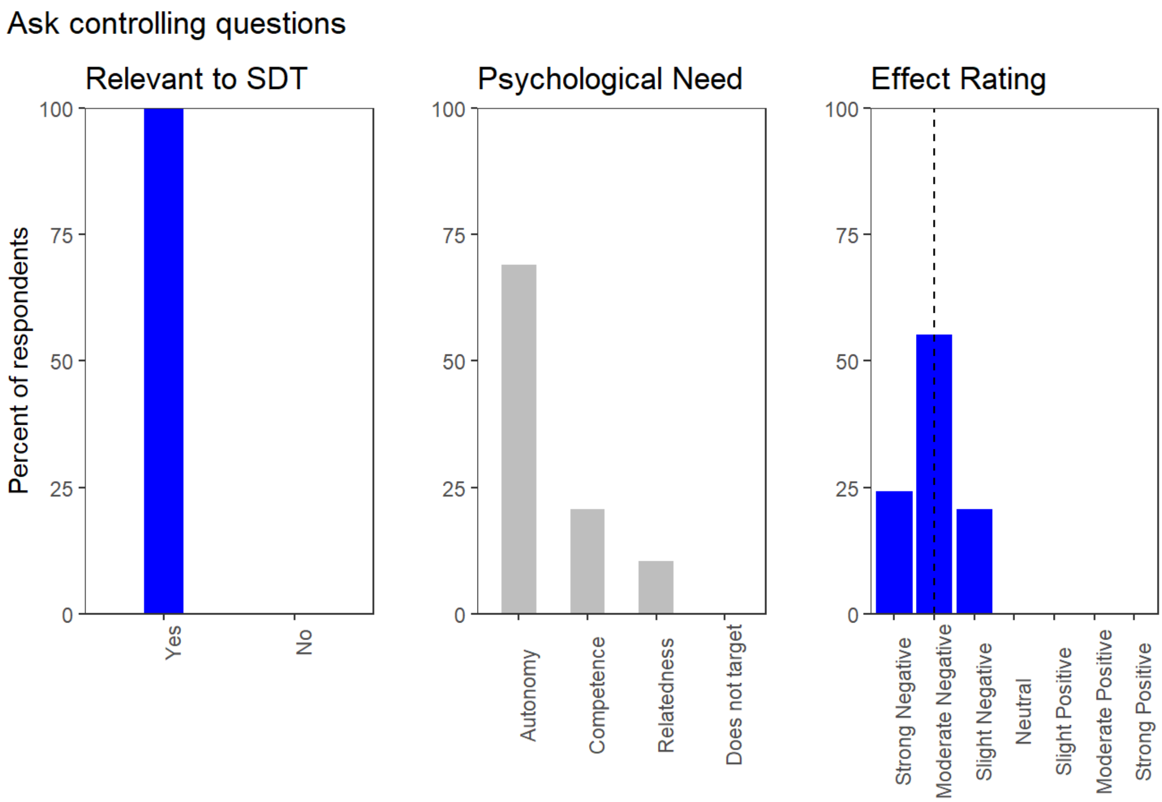
Provide commands that are phrased as rhetorical questions

Example Behaviour:

"Can you just do it like I showed you?”

Function Description:

Communicates disapproval for the students current behaviour without clarifying how to improve or a rationale for change.



***

TMB#54

Deadline statements

Description:

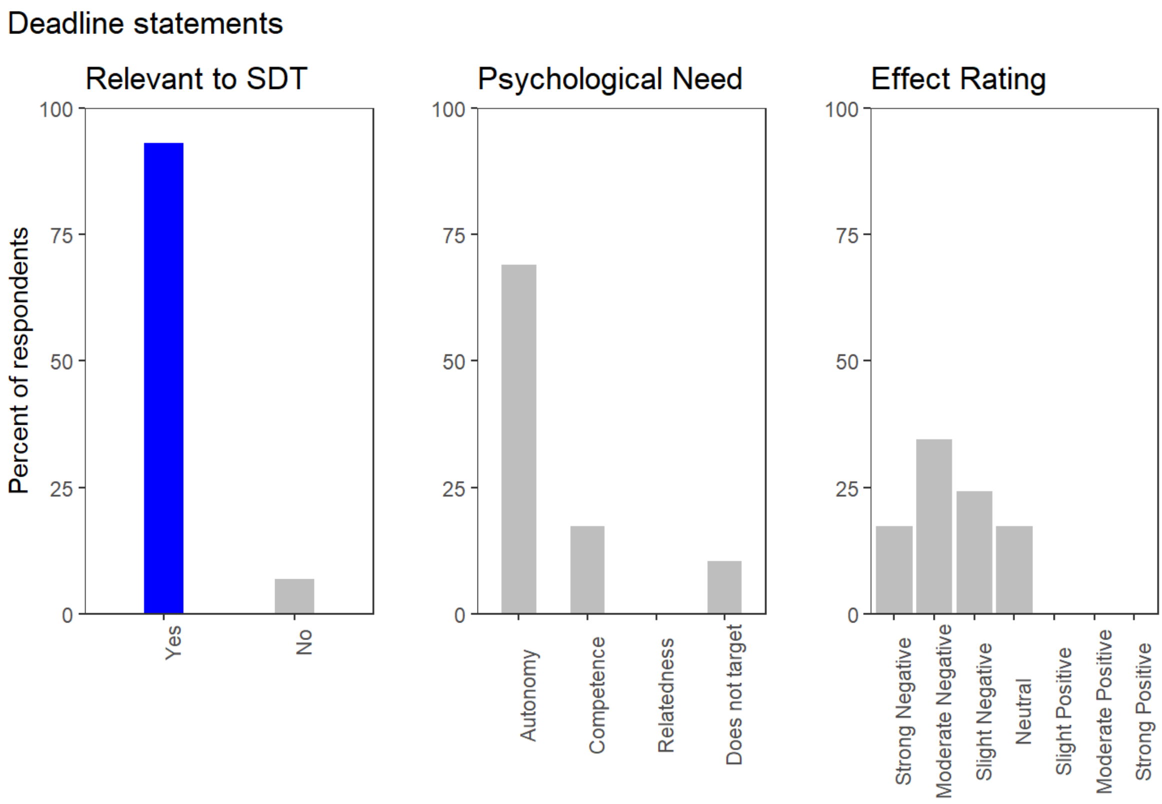
Tell the class they are running out of time

Example Behaviour:

"A couple of minutes left"; "We only have a few minutes left"

Function Description:

Adds pressure on students to work faster and finish tasks for extrinsic reasons



***

TMB#55

Outlining Punishment Contingencies

Description:

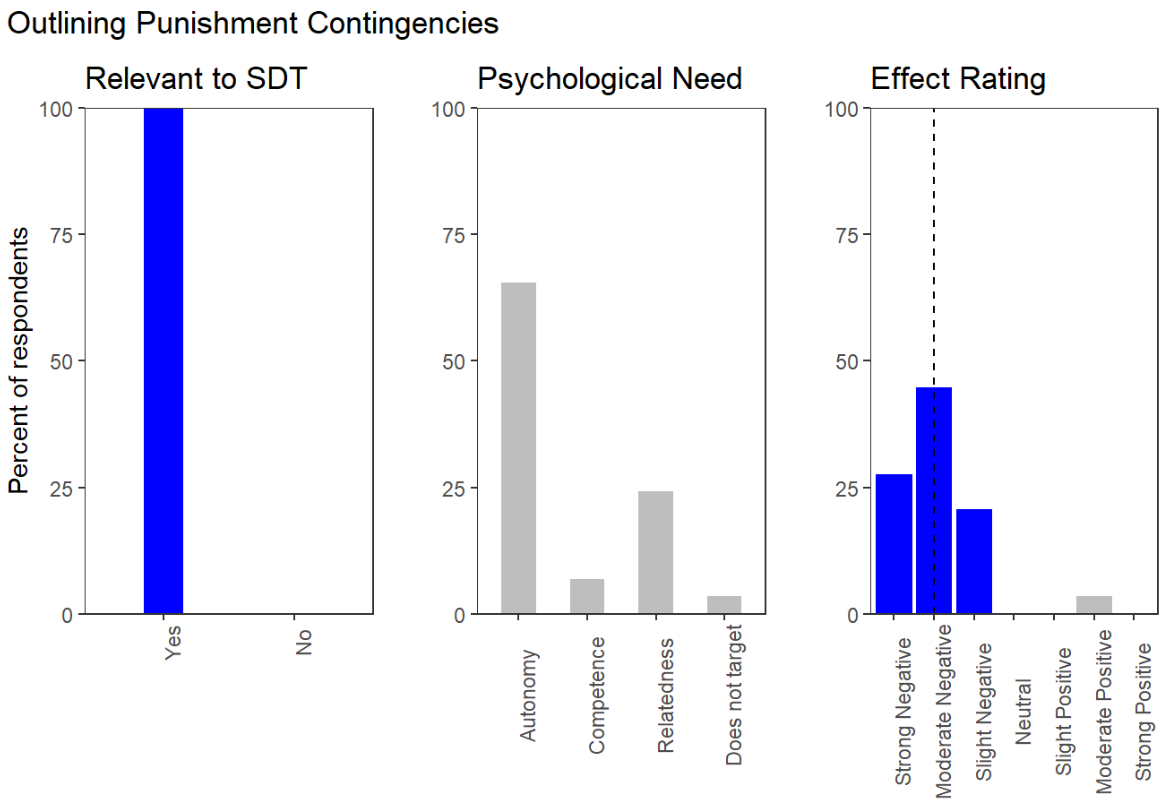
Declaring (but not yet enforcing) if-then extrinsic punishments—contingencies that are not inherent to the task and are provided in an effort to extinguish a behaviour

Example Behaviour:

"If you two speak one more time, I will send you out"

Function Description:

Imposes an extrinsic reason for student behaviour.



***

TMB#56

Provide fair punishments

Description:

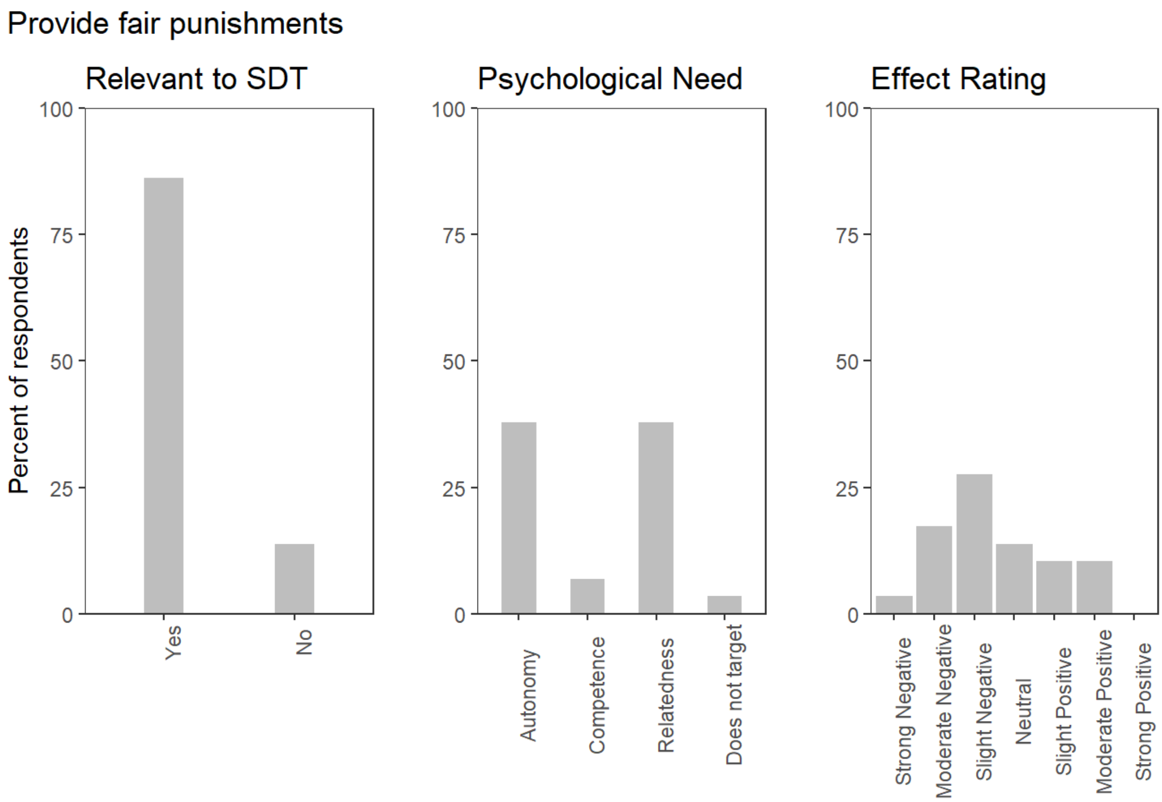
Provide punishments fairly so students who misbehave are treated equally

Example Behaviour:

Sending both of two students out of class when they misbehave or break a rule

Function Description:

Ensures misbehaviour is consistently and reliably met with external contingencies



***

TMB#57

Providing Rewards

Description:

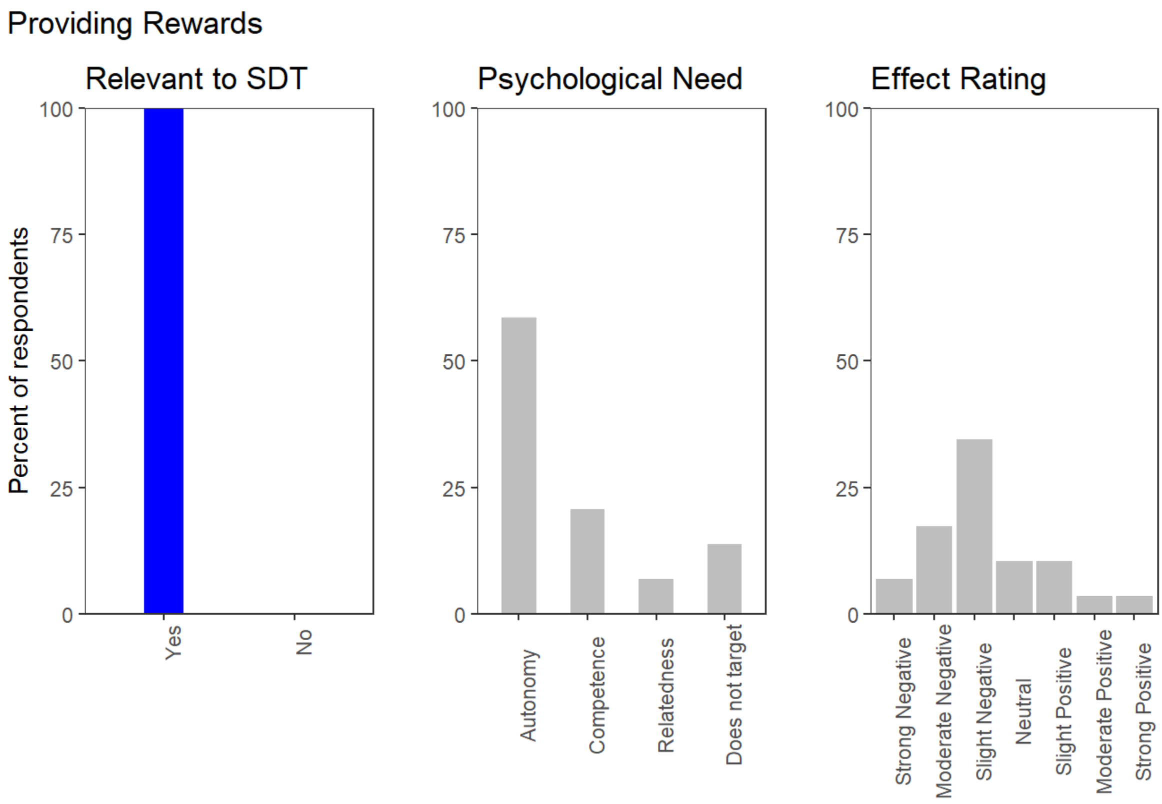
Provide rewards when the expected behaviour is observed

Example Behaviour:

"You all did your homework, so as I promised, we can watch a YouTube video today"

Function Description:

Adds external, tangible signal of which behaviours are desirable/valued by the teacher



***

TMB#58

Conditional positive regard

Description:

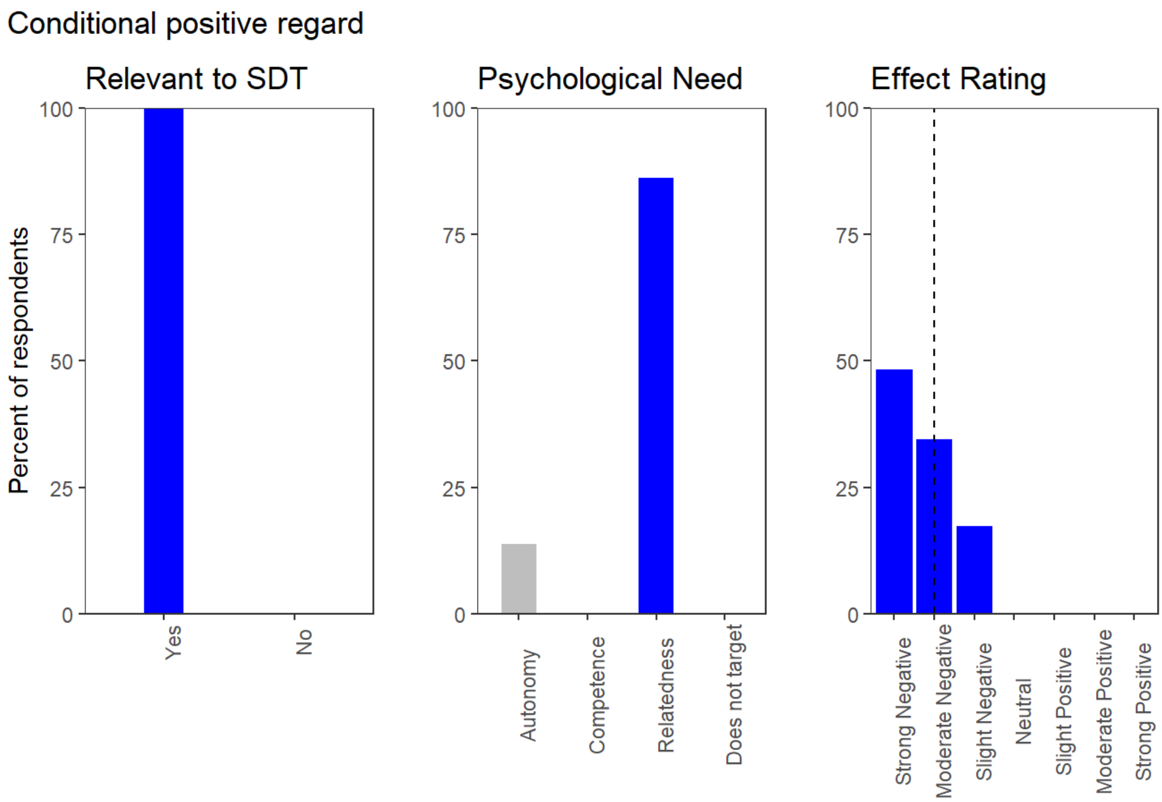
Withdrawal warmth from a student in response to poor behaviour

Example Behaviour:

"I am sick of your behaviour"; disdainfully glaring at student

Function Description:

Demonstrate that attention and warmth are contingent upon meeting the teachers expectations of good student behaviour



***

TMB#59

Abusive language (content)

Description:

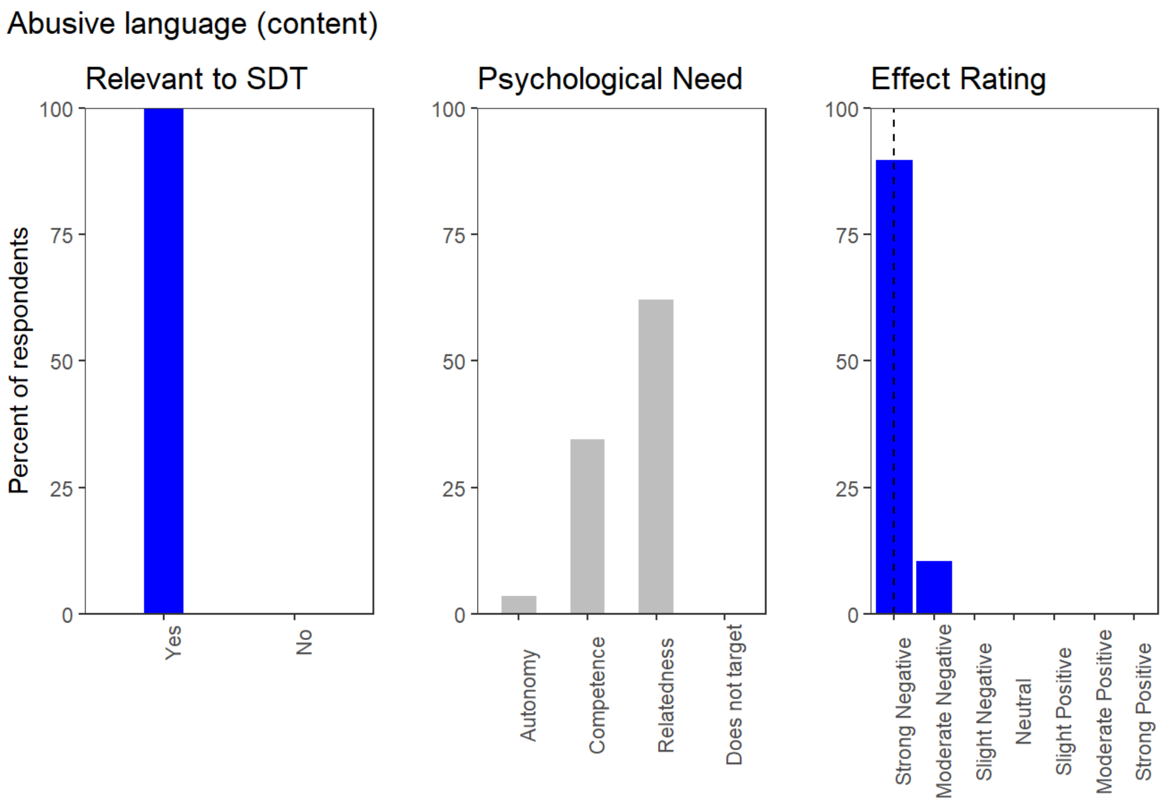
Calling students by hurtful names when they misbehave

Example Behaviour:

Calling a student "dummie" or "moron"

Function Description:

Performance mistakes and behavioural misconduct are met with competence-threatening punishment



***

TMB#60

Yelling or harsh tone

Description:

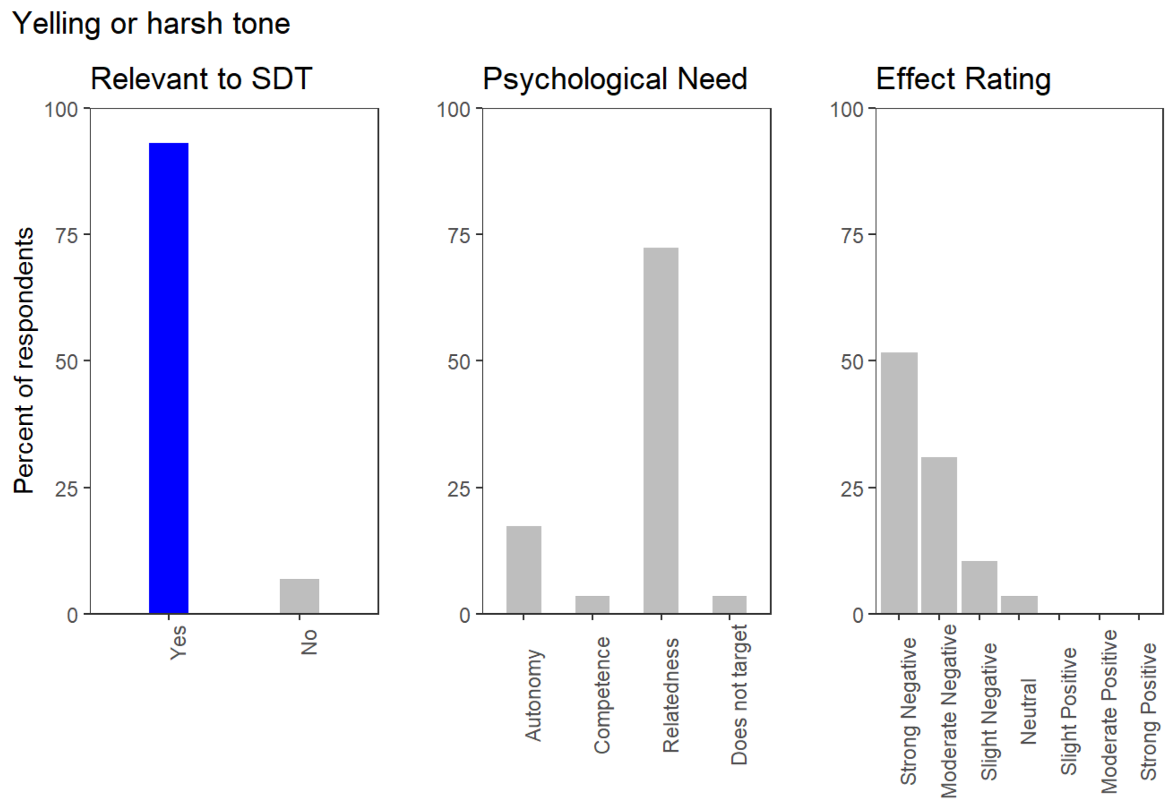
Teacher yells to get control of the classs

Example Behaviour:

Yelling such as "HEY!"; "STOP IT!"

Function Description:

Creates a more emotionally unstable and unpredictable environment for students, increasing fear



***

TMB#61

Unfair rewards

Description:

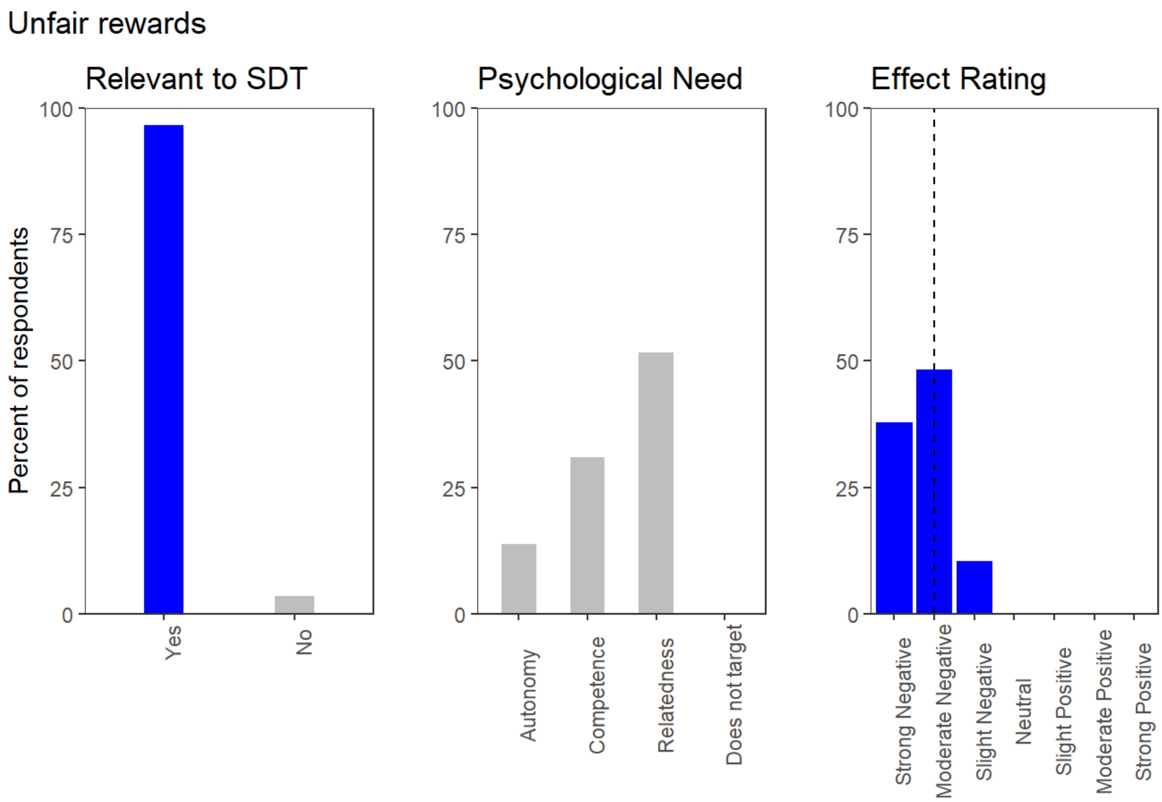
Provide rewards unfairly so students who are doing equally well, get different rewards

Example Behaviour:

Rewarding only one of three people who all completed a task

Function Description:

Students feel rewards are not predictable and teacher behaviour unjust



***

TMB#62

Unfair punishments

Description:

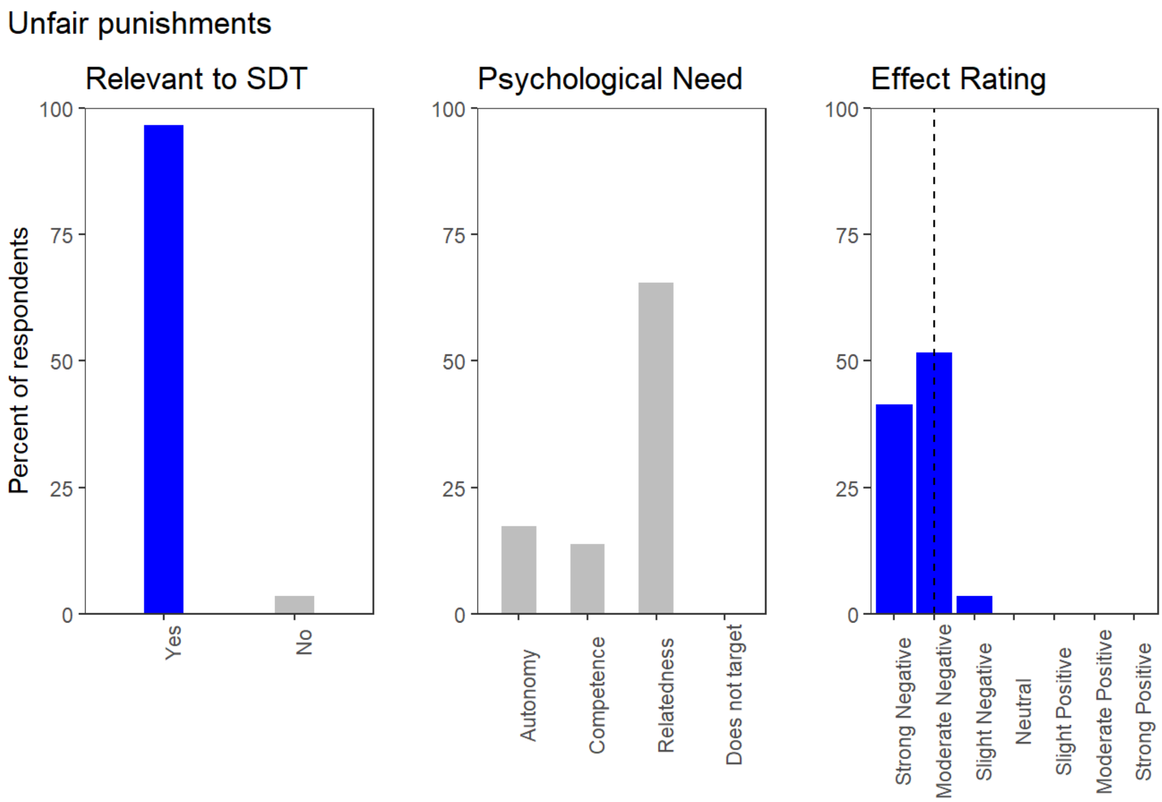
Provide punishments unfairly so students who misbehave are treated unequally

Example Behaviour:

Punishing only one of two students who are speaking out of turn

Function Description:

Means structures are perceived as inconsistent and unreliable



***

TMB#63

Criticism / fixed quality

Description:

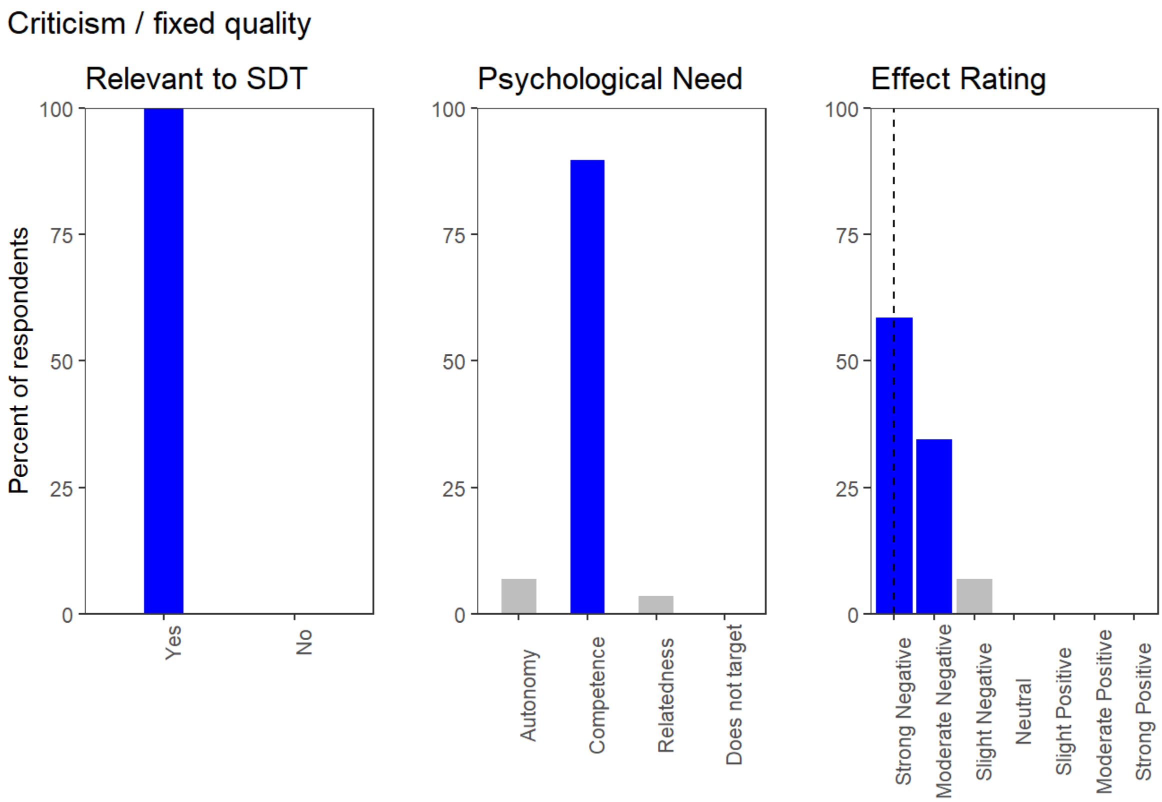
Provides criticism that targets a fixed quality

Example Behaviour:

"you are not tall enough", "maths is not your strength"

Function Description:

Emphasises the importance of inherent (e.g., genetic) abilities for achieving success



***

TMB#64

Criticism / losing or peer comparison

Description:

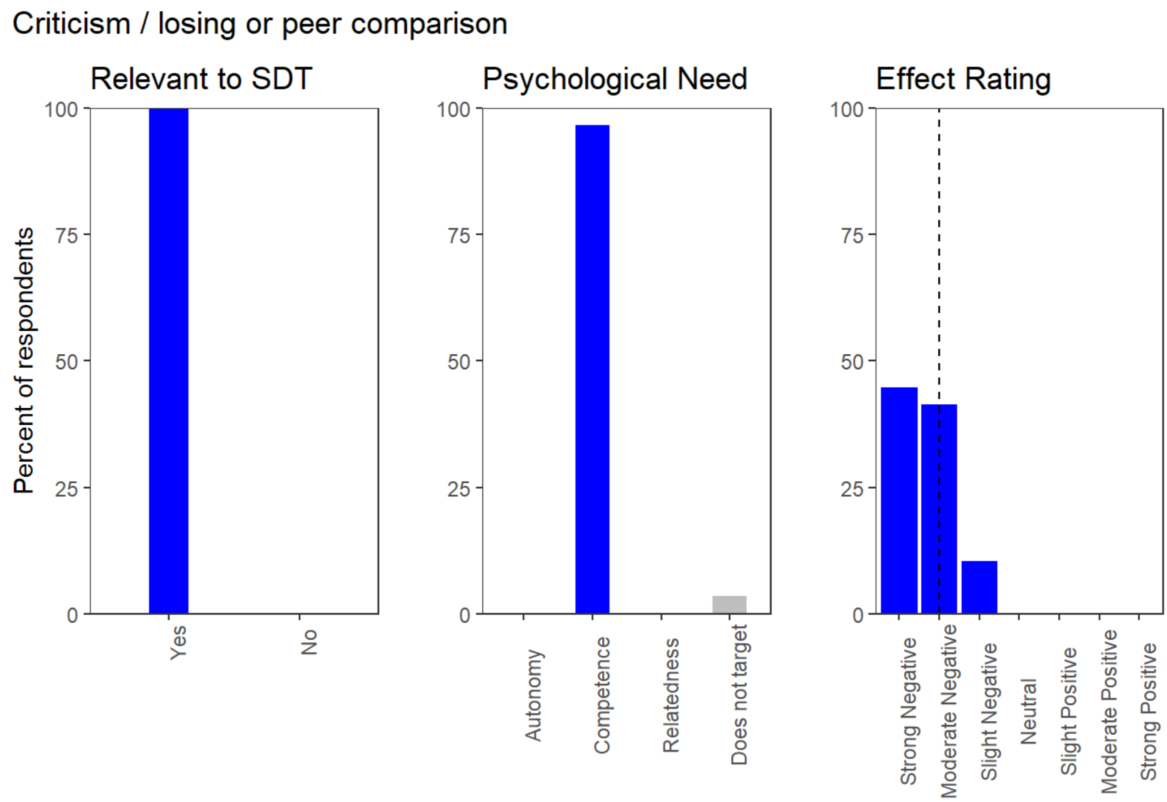
Tell students when they are not doing as well as others

Example Behaviour:

"You could learn from Paula who is better at this"

Function Description:

Emphasises peer comparison for establishing a sense of competence, meaning few students experience success by being the best



***

TMB#65

Criticism / public

Description:

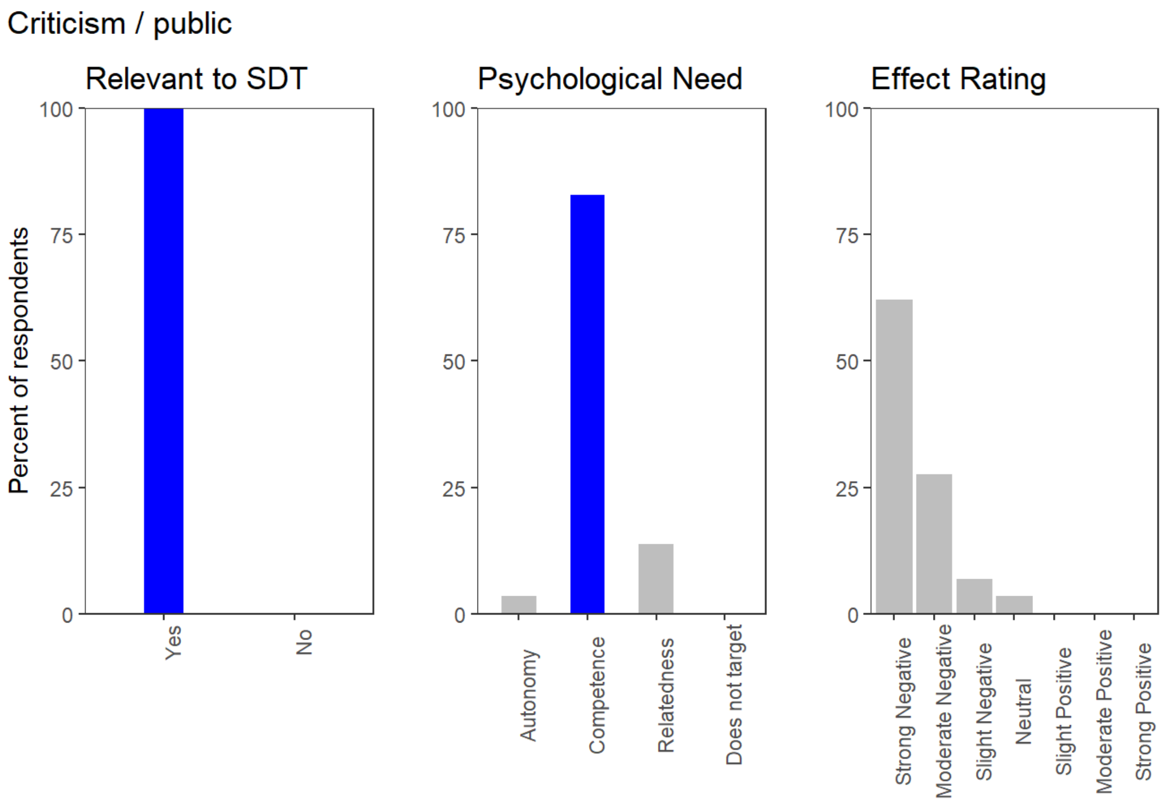
Provide feedback in public so other students can hear

Example Behaviour:

Provide feedback in front of the class

Function Description:

Increases risk of feedback being ego-threatening



***

TMB#66

Criticism / Unclear (Vague)

Description:

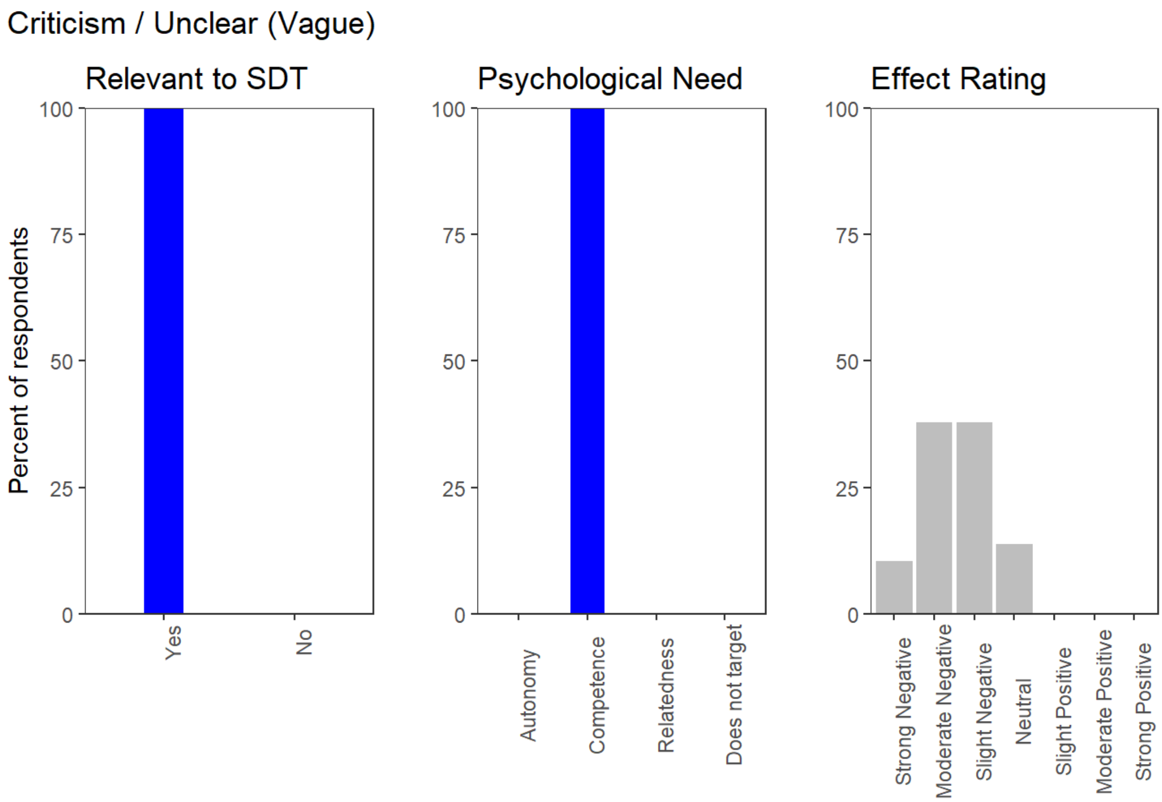
Provides vague criticism with no instruction of how to improve

Example Behaviour:

"Come on, James, you need to do better"

Function Description:

Creates ambiguity regarding strategies for students to increase competence



***

TMB#67

Exhibiting or uttering solutions / answers

Description:

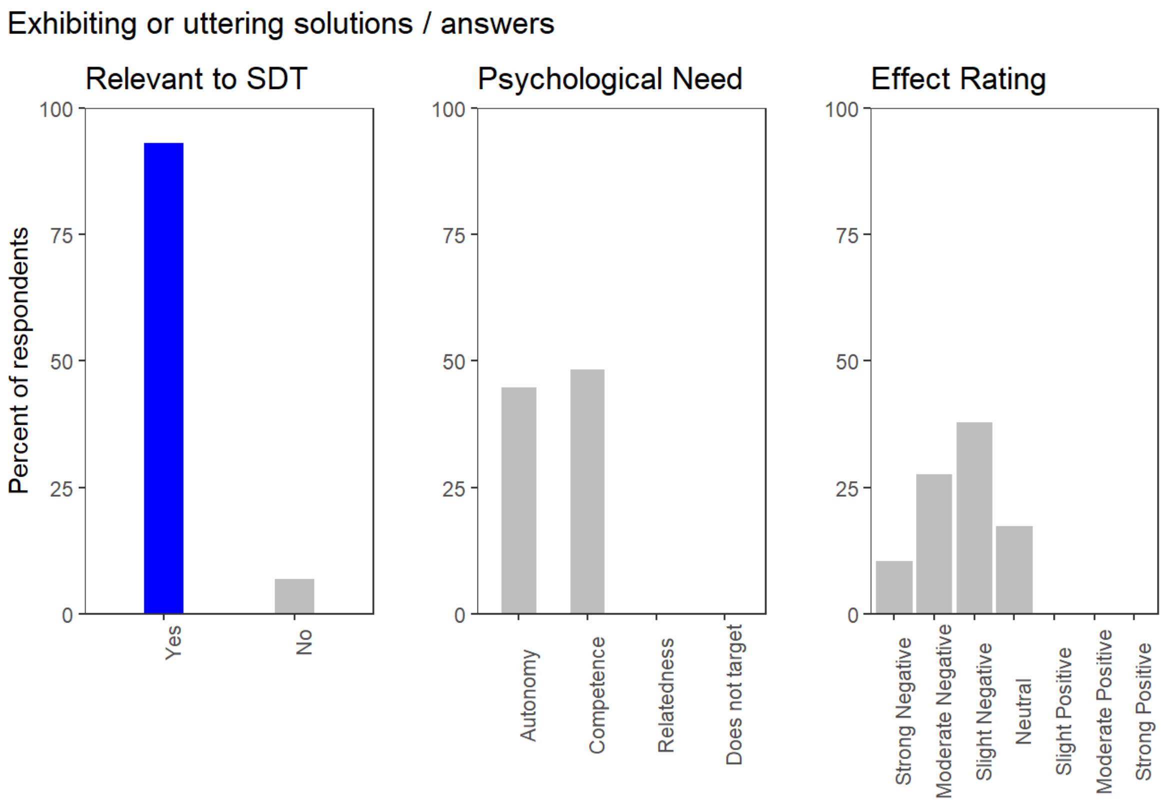
Give answers to problems instead of letting students figure it out

Example Behaviour:

"The answer is 42"

Function Description:

Stifles self-directed learning and provides external locus of causality for succes (i.e., from the teacher)



***

TMB#68

Praise / winning or peer comparison

Description:

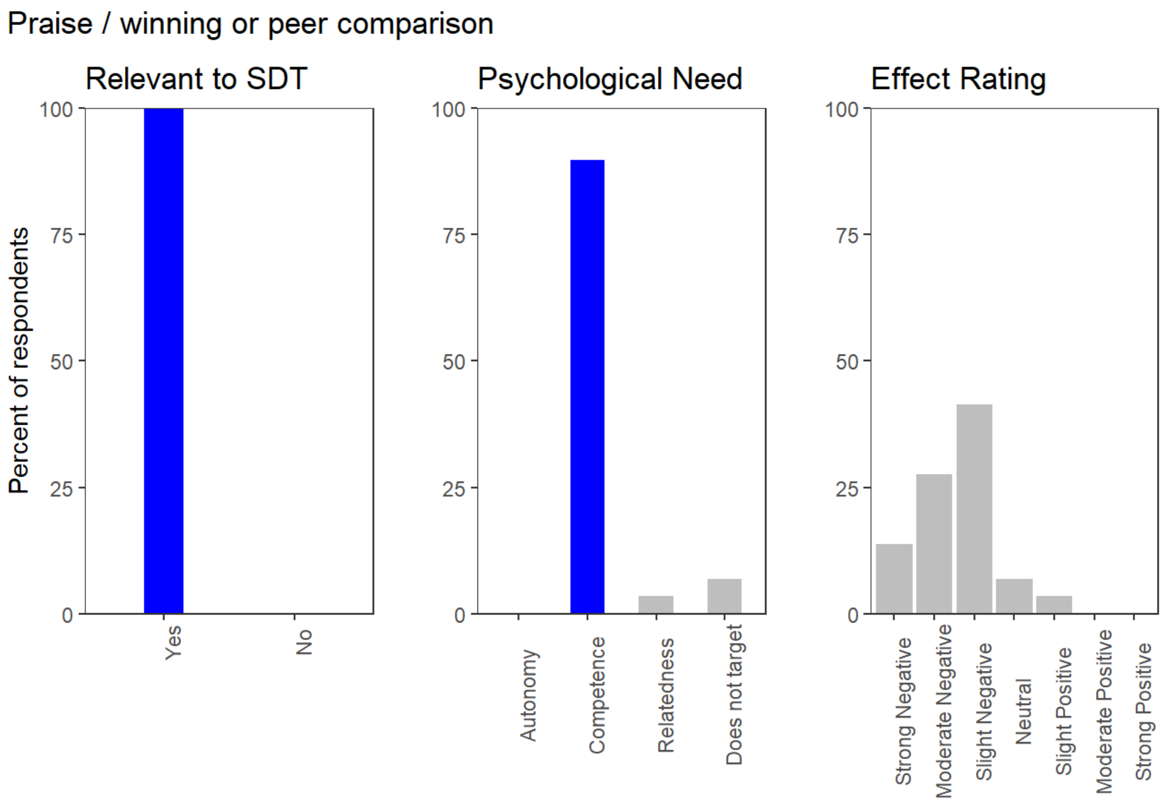
Congratulate winners so that everyone knows who did the best

Example Behaviour:

"The highest score on the exam was John"

Function Description:

Emphasises peer comparison and establishing a sense of competence, meaning few students experience success by being the best



***

TMB#69

Praise as contingent reward

Description:

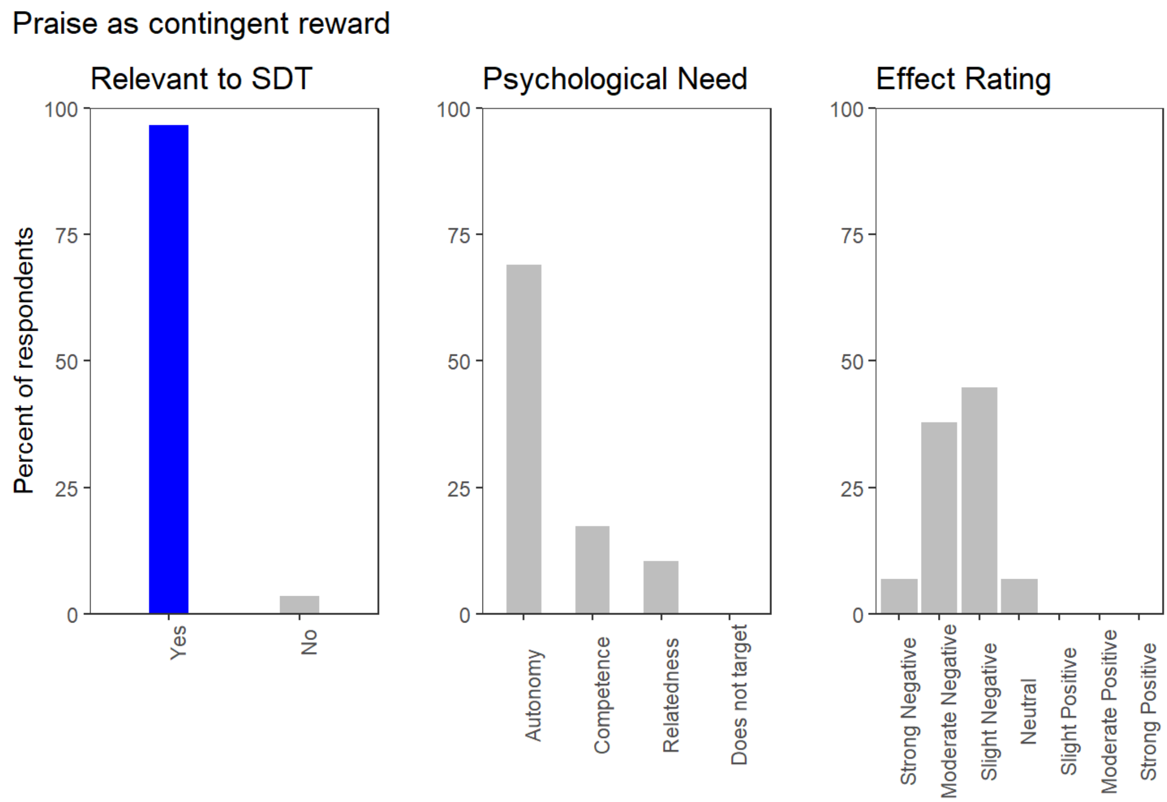
Praise students only when they do what they are told

Example Behaviour:

Teacher says to a student "You are smart" when they do what they were told

Function Description:

Increases perceived external incentives for doing an activity that is favoured by a teacher



***

TMB#70

Exclusionary activities

Description:

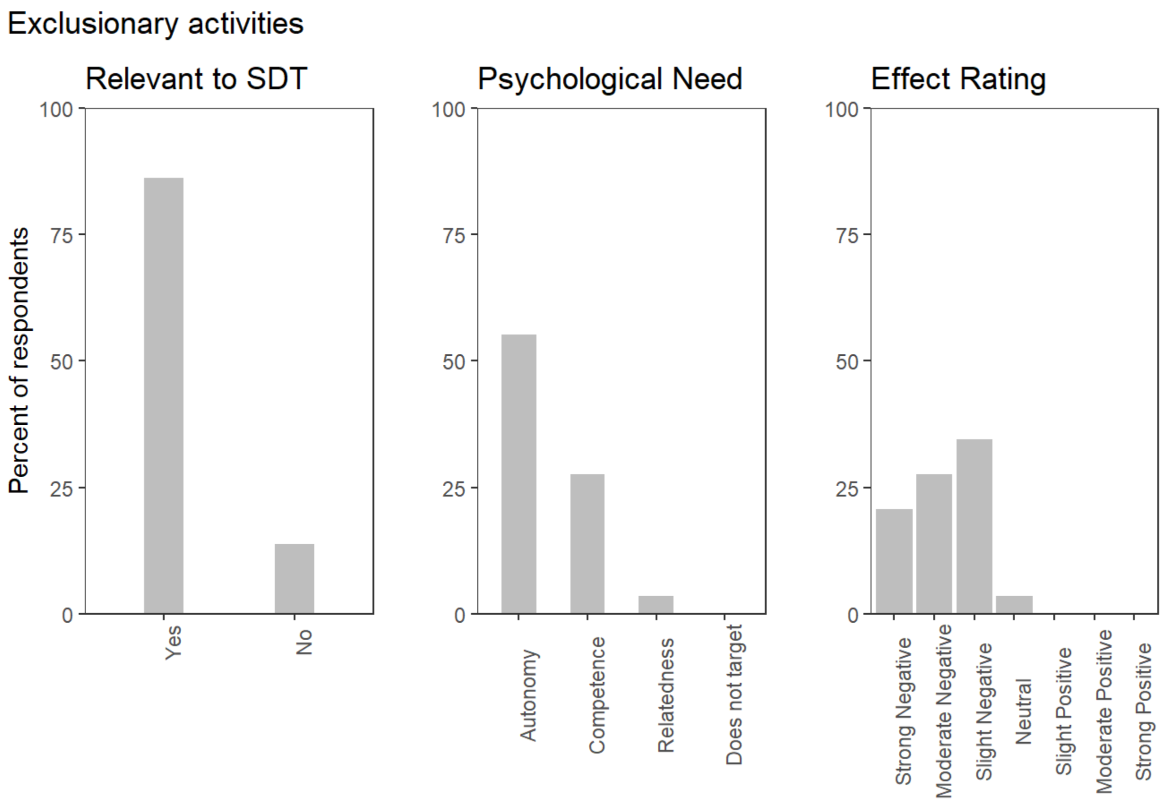
Set up activities so there are times where some students are not doing anything

Example Behaviour:

"if you have finished the questions, just sit quietly until everyone else is finished"

Function Description:

Students do not have opportunities to engage, even if they want to



***

TMB#71

Goals / Competition

Description:

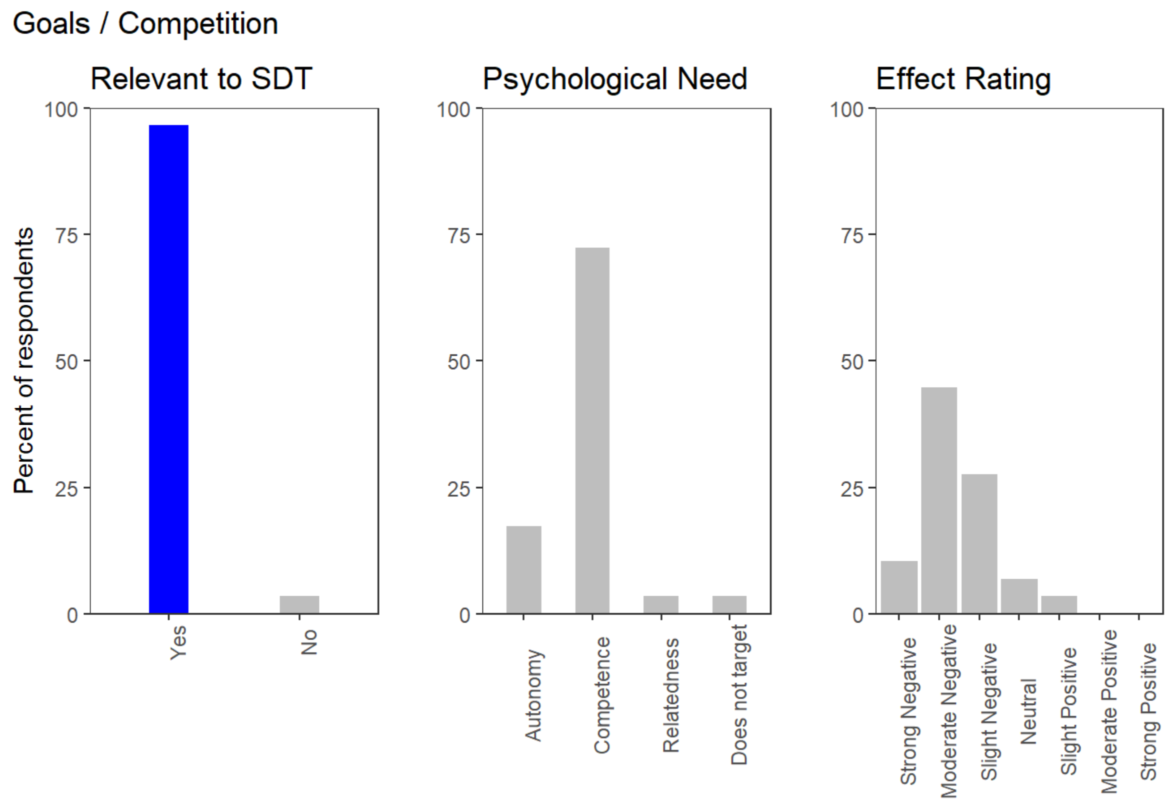
Set up activities where the goal is to do better than other students so that students compete against each other

Example Behaviour:

"whoever completes these problems in the fastest time wins"

Function Description:

Provides extrinsic reasons for working hard and few opportunities for success (i.e., winning)



***

TMB#72

Homogeneous Groups

Description:

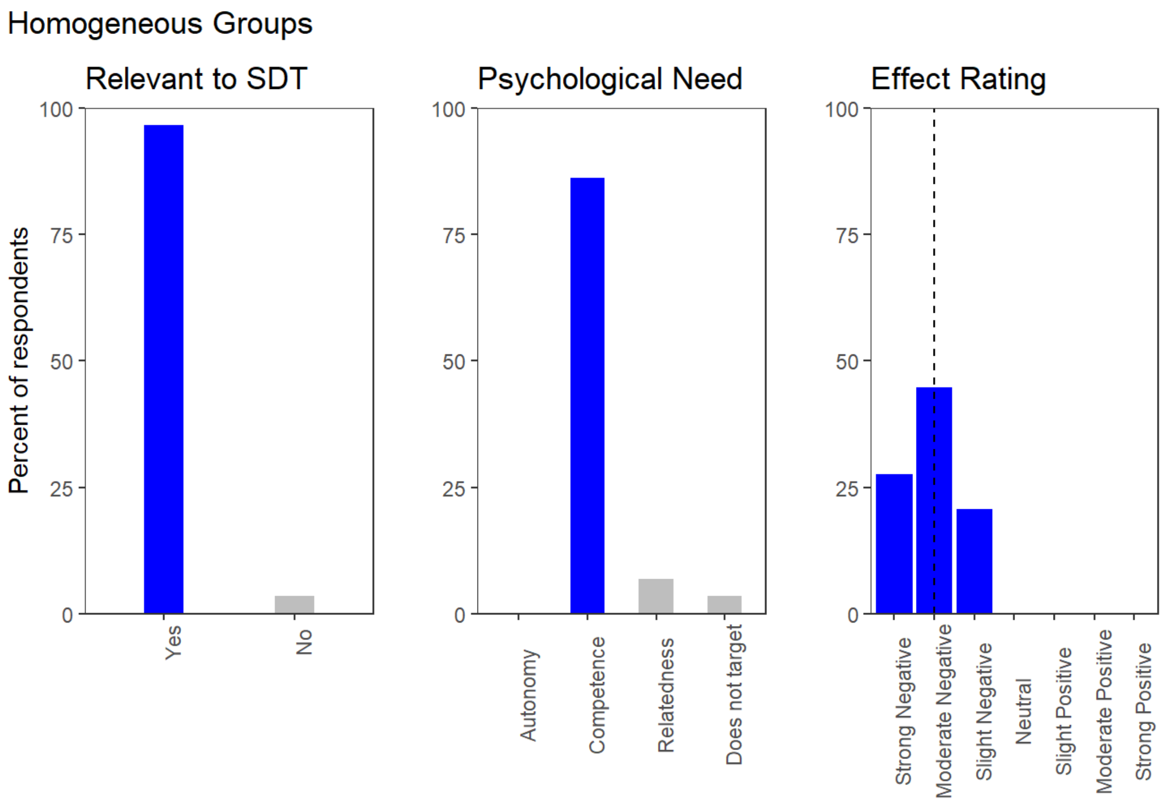
Grouping is done publicly and students are put in groups based on their ability so that there are "top" and "bottom" groups

Example Behaviour:

"if you got more than 7/10, join this group. Less than 7: in this group. If you did not do your homework, you are at the back"

Function Description:

Increases public signalling of student incompetence



***

TMB#73

Chaotic Teaching

Description:

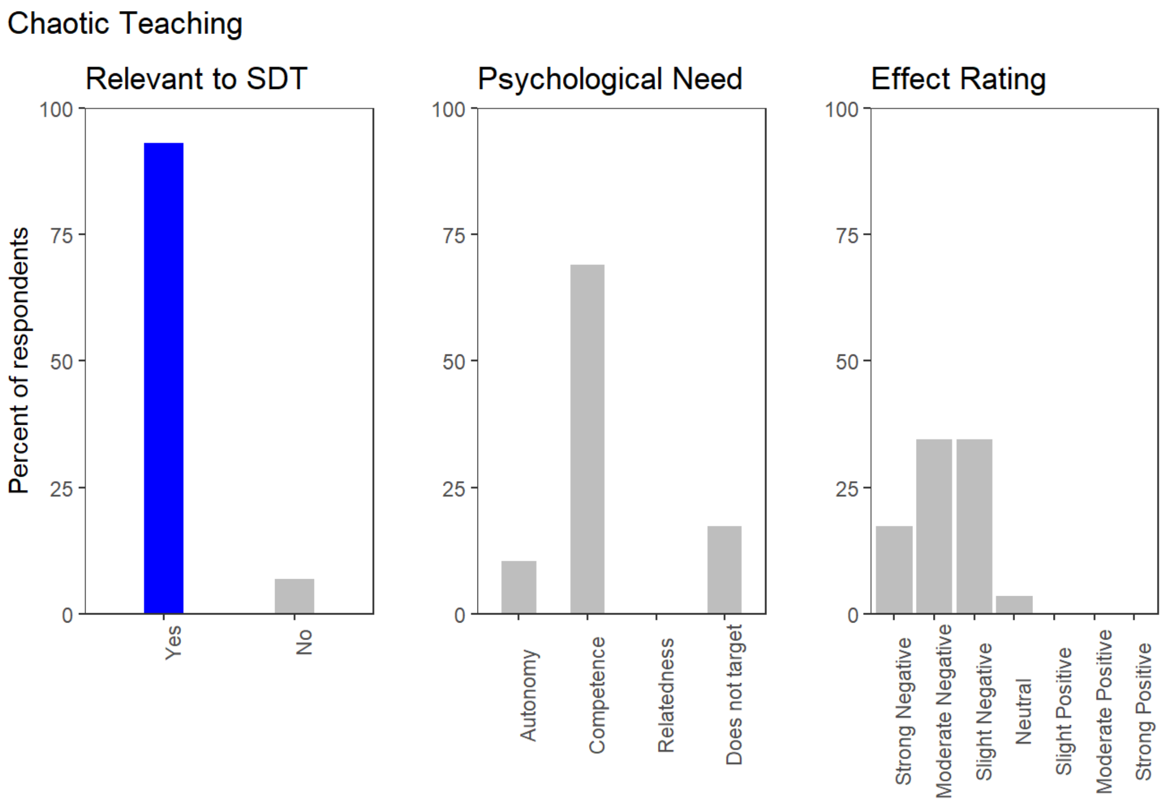
Leave students without clear instructions so the class waits while the teacher does something else

Example Behaviour:

Teacher leaves students waiting when arranging papers at front

Function Description:

Students do not know what they should be doing to learn



The following TMBs are suggested by the panellists.

New TMBs suggested by the experts

TMB#74

Teaching children to set intrinsic life goals for learning

Description:

Teach students to set intrinsic life goals for learning such as "overtly healthy attitudes toward the learning process (e.g., embracing challenges, enjoyment of learning), "helping others" (e.g., how it applies to helping others or bettering one's community).

Example Behaviour:

A teacher might explain to the class what previous students have found interesting about a homework assignment (e.g., "Reading helps me to gain knowledge about life") or how the underlying skills and knowledge can prepare the student to help others ("I want to use my reading skills to read to little kids").

Function Description:

increases children's autonomous motivation to learn and enjoyment of learning.\nOR\nStudents will try to understand the lessons more, become better at doing the activities, so that students can help others someday, or discover something interesting.

***

TMB#75

Use parables, stories, analogies, or metaphors

Description:

Use parables, stories, analogies, or metaphors to help students to connect abstract constructs into concrete examples

Example Behaviour:

"mistakes are stepping stones, not stumbling blocks"; "treating others well is like sowing good seeds, eventually you'll reap a good harvest"

Function Description:

Promote empathy through narratives that students can connect to, and provide examples that students could follow.

***

TMB#76

Adopting student initiatives

Description:

Take student suggestions into learning activities when they arise

Example Behaviour:

"That's a great idea. We can do that activity in this session."

Function Description:

Encourages and rewards student initiative and self-management of learning.

***

TMB#77

Set goals on behalf of the class

Description:

Teachers setting expectations for students rather than letting them to decide their own goals

Example Behaviour:

"The goal for tonight is to complete the all activities on page 4."

Function Description:

Provides extrinsic reasons for the goals and for doing activities and the student autonomy is undermined

***

TMB#79

Group students with similar interests

Description:

Create groups in the class where students with similar values or interests can work together on problems

Example Behaviour:

When studying geography, grouping musical students to look at a country's music, the sporty students to look at the country's sports, and other students to look at the country's key historical events.

Function Description:

Allows students to work on tasks—and with people—that match their interests and values.

***

TMB#80

Regular communication with parents

Description:

Teachers engaging in regular contact (e.g., phone, email, text) with parents about the activities in the class or of their children

Example Behaviour:

The teacher calls a parent when she notices that a student has been particularly disengaged and unenthusiastic to talk

Function Description:

Supports connections between the students' home and school life, identifying ways that key people in each domain (e.g., parents, teachers) can support each-other

***

TMB#81

Be sarcastic

Description:

Use sarcastic negative phrases

Example Behaviour:

"Class started 3 minutes ago. Soooo nice of you to join us." Or, "It's not like what we are learning today is important or anything."

Function Description:

Reduces student self-esteem and devaluates their sense of being a valuable person and students

***

TMB#82

Unfair use of praise

Description:

Praises students unfairly or unequally; shows favourites

Example Behaviour:

Complementing only one of three students who completed a problem a creative way

Function Description:

Makes students feel like some are more worthy of praise than others

Appendix C.2

Delphi Round 2 Results with Plots

TMB#1

Active Learning

Description:

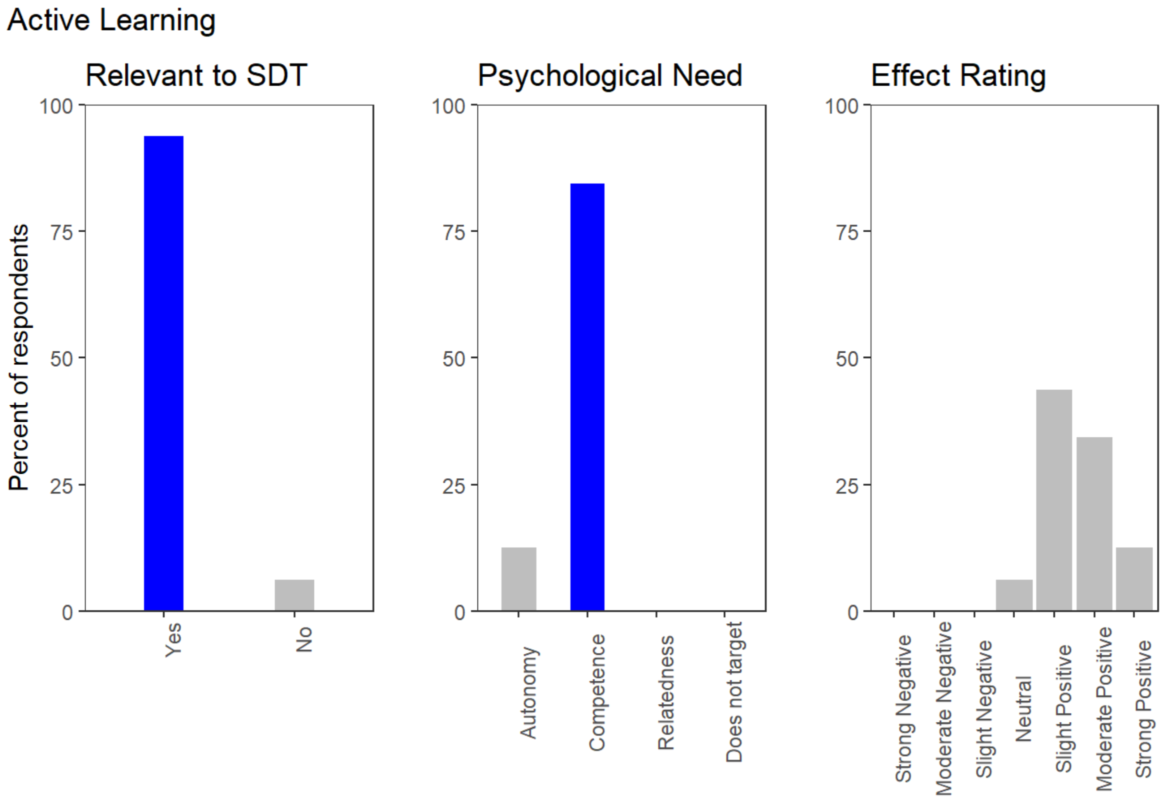
Set up activities where all students are engaged in a learning activity or solving a problem

Example Behaviour:

"Use these numbers to figure out how heavy the Sydney Harbour Bridge is"

Function Description:

Allows each student hands-on practice with an activity designed to progress development of a skill



TMB#2

Demonstrating examples

Description:

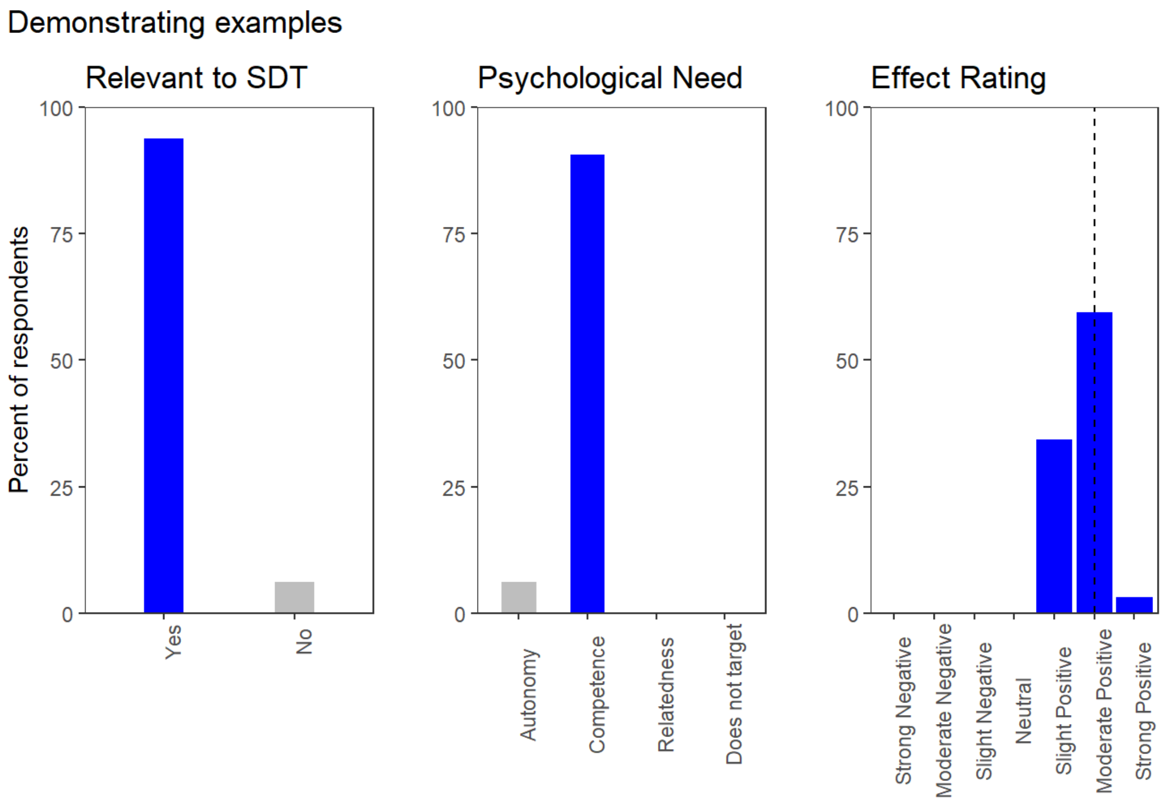
Modelling or demonstrating examples

Example Behaviour:

When throwing, see how my other hand points at the target?

Function Description:

Provides template for student to follow



TMB#3

Discuss class values

Description:

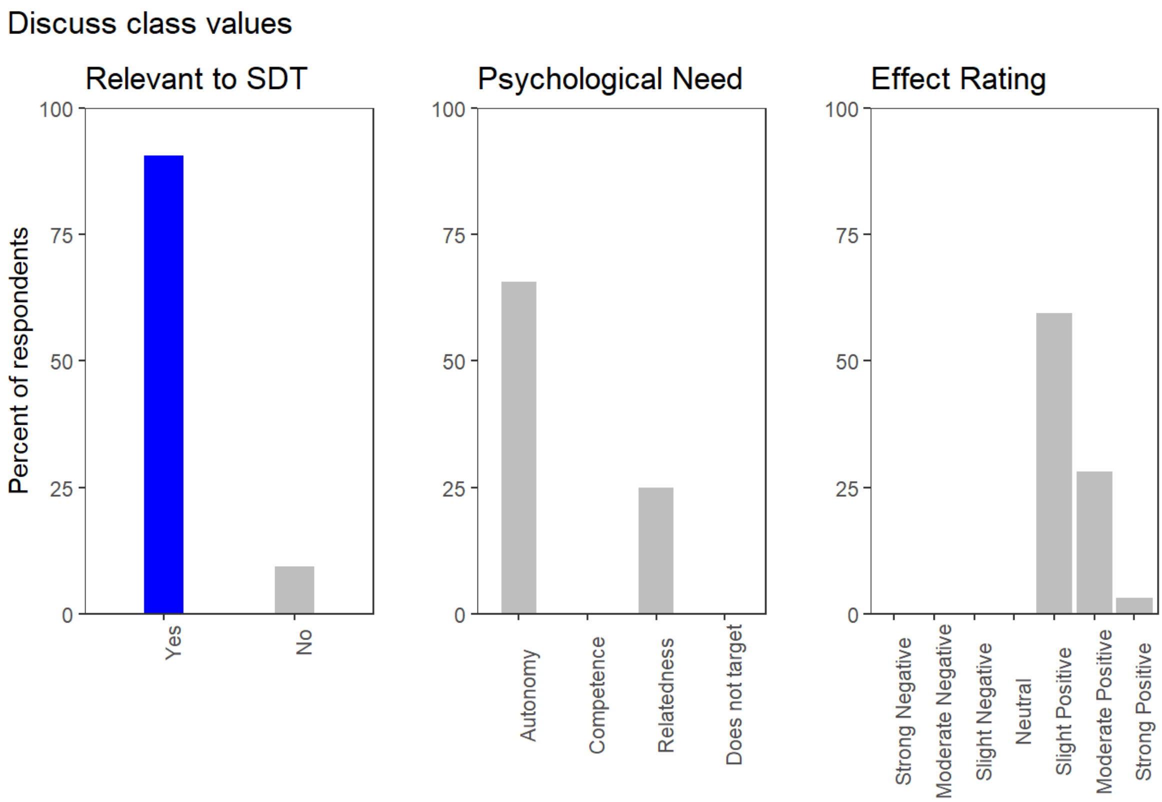
Collaboratively establish the values important to display in the class, or remind students of the collaboratively derived values

Example Behaviour:

"We all thought working hard was important, so even though many find this task difficult, see if you can push through to the end."

Function Description:

Connects the activities that take place in class with values that the student cares about



TMB#4

Modelling resilience by expressing vulnerability

Description:

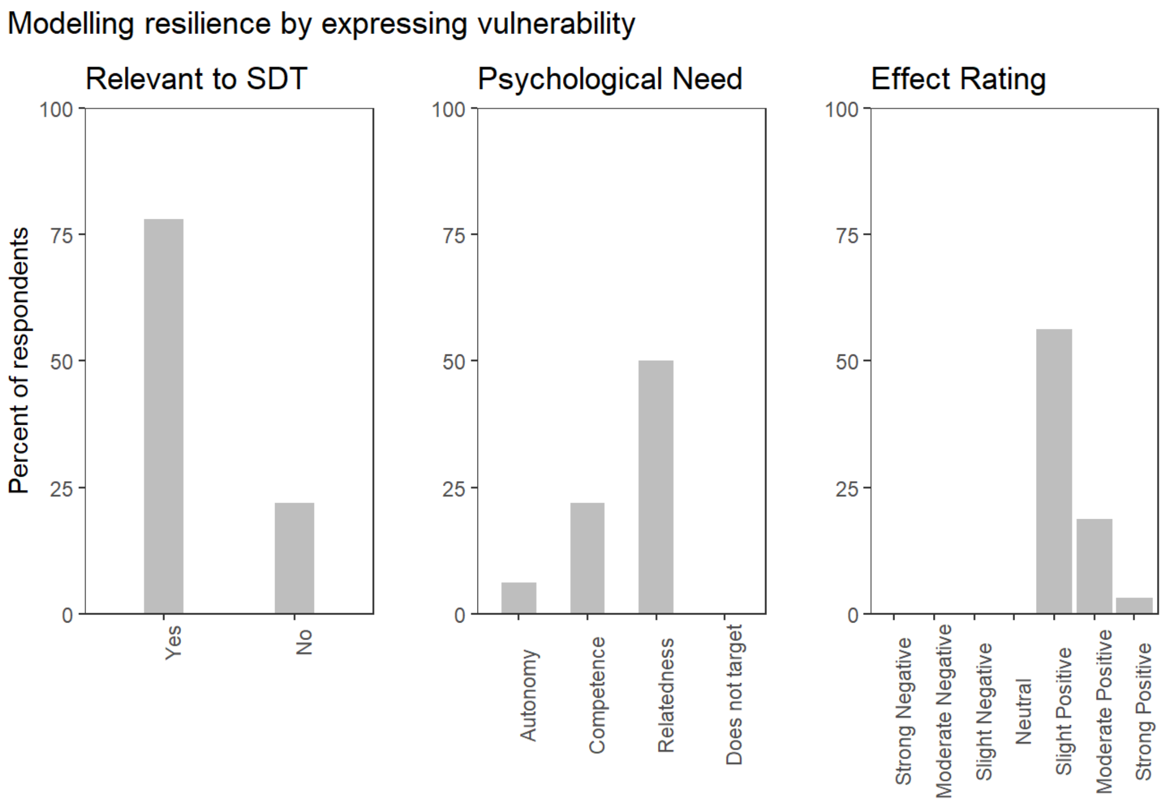
Showing that it is possible to adapt and achieve despite difficulties now or in the past

Example Behaviour:

"I struggled to write clearly for years, but I kept asking for feedback and got better."

Function Description:

Helps students to perceive the teacher as a model for coping with challenges



TMB#5

Teacher enthusiasm

Description:

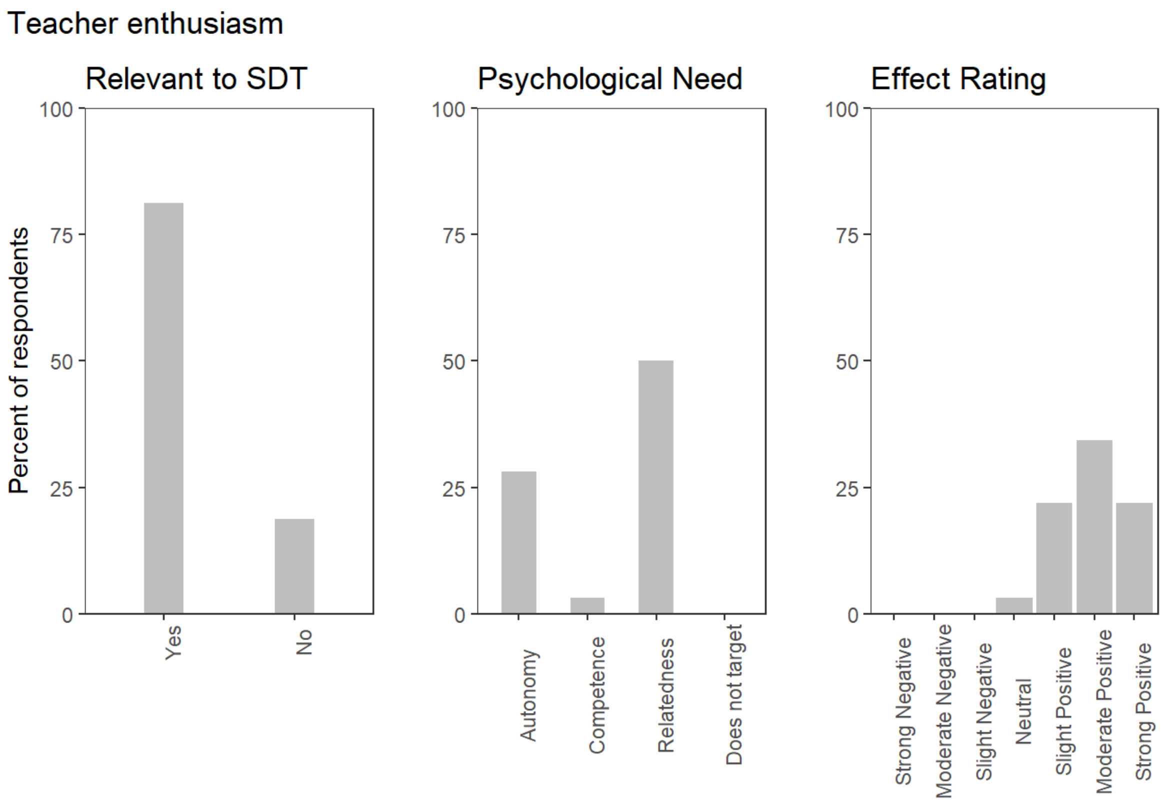
Present content enthusiastically to make things fun and interesting

Example Behaviour:

"Now we are going to learn to something really interesting!"

Function Description:

Models the attitude and energy that the teacher would like the students to demonstrate; shows interest in the material.



TMB#6

Offer rewards

Description:

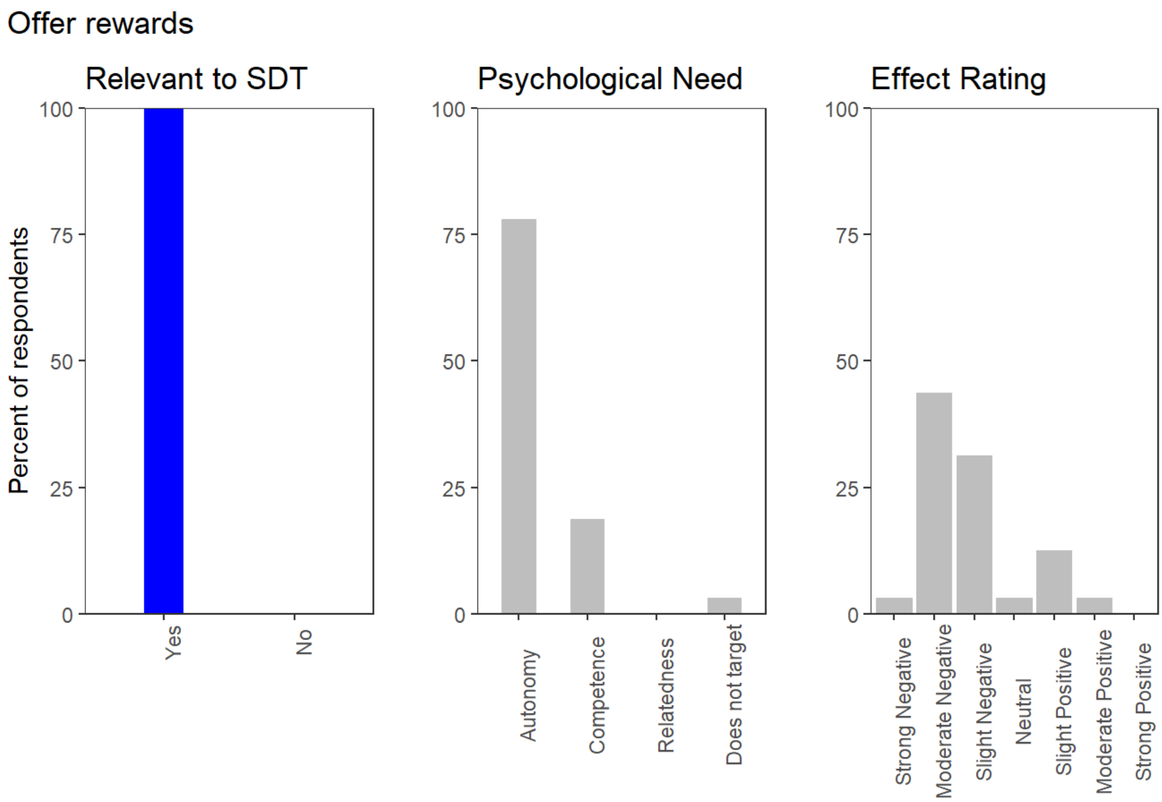
Offering—but not yet providing—extrinsic rewards: privileges or items that are not inherent to the task, but are provided in an effort to promote a behaviour.

Example Behaviour:

"If you all finish the questions, I'll play a short video clip."

Function Description:

To direct behaviour so students know what behaviour the teacher wants to see



TMB#7

Fair use of praise

Description:

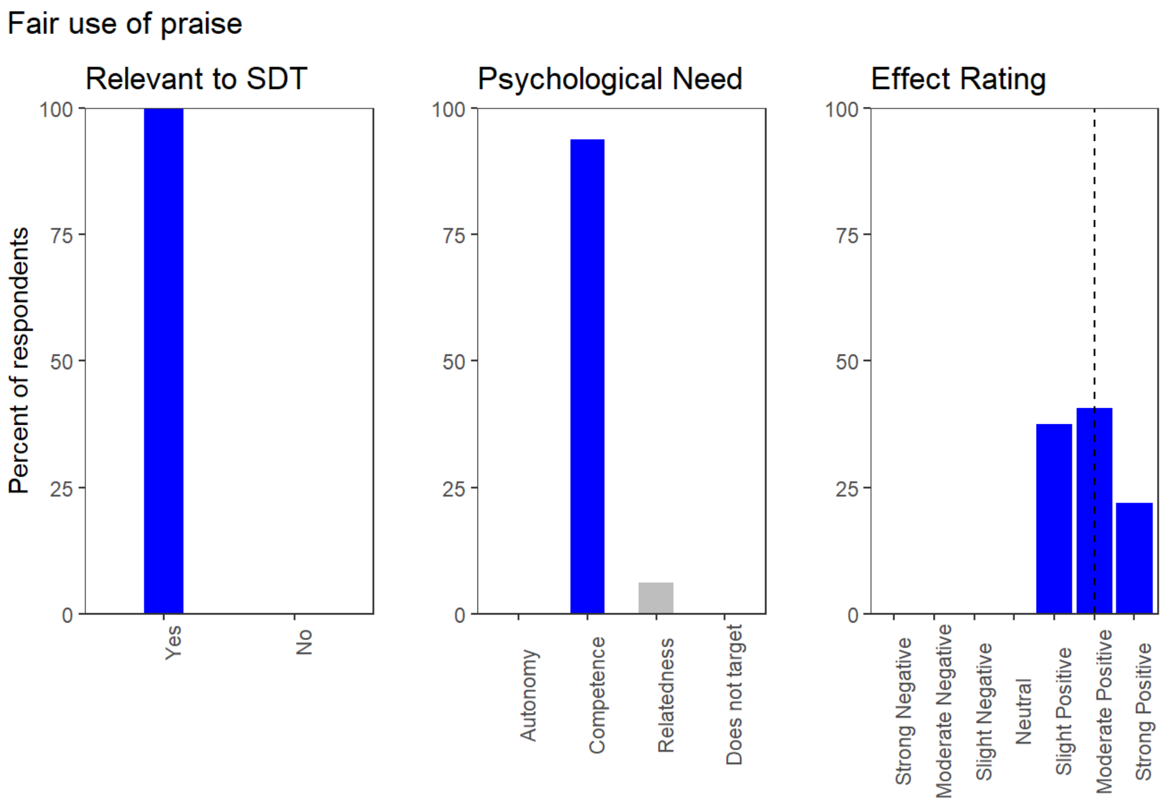
Appraises a student to help him/her improve or increase effort

Example Behaviour:

Complementing all three people who completed a project in specific ways

Function Description:

Increases sense of efficacy, inclusion and belonging



TMB#8

Provide Feedback in Private

Description:

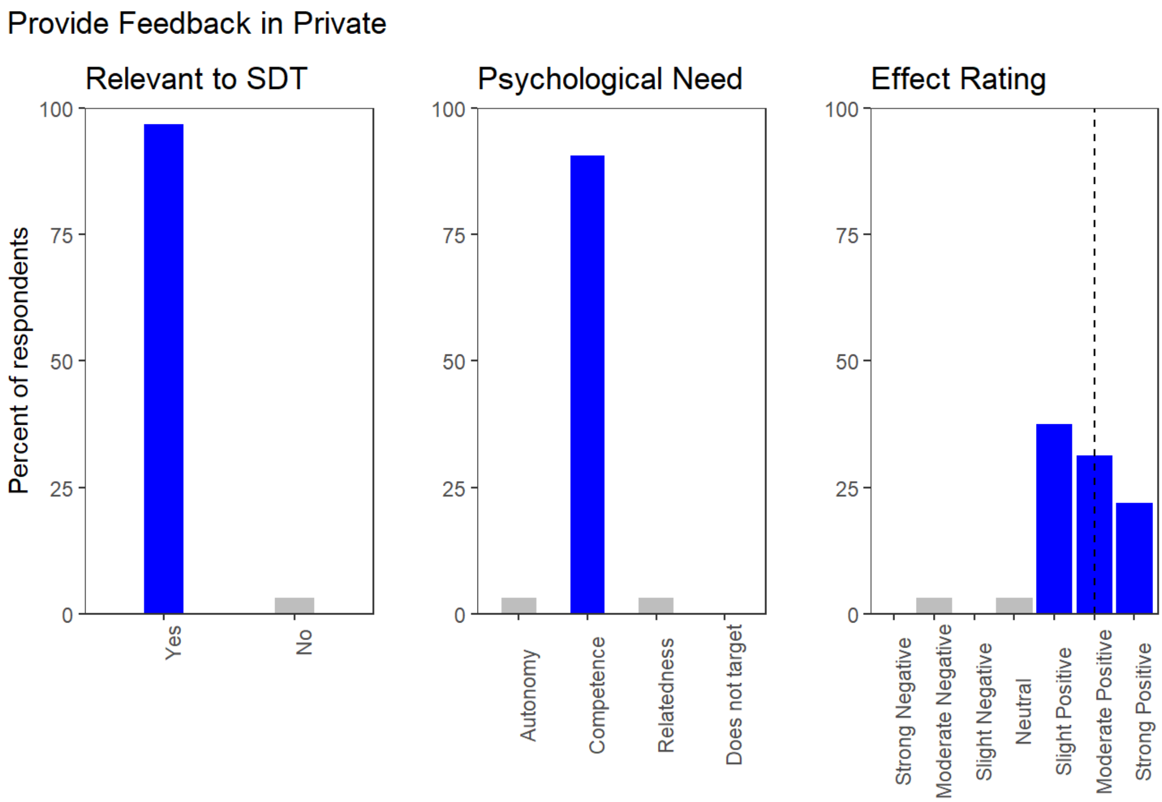
Provide any kind of sensitive or critical feedback in private

Example Behaviour:

Provide feedback 1 on 1 with the student

Function Description:

Mitigates risk of feedback being ego-threatening



TMB#9

Provide specific feedback

Description:

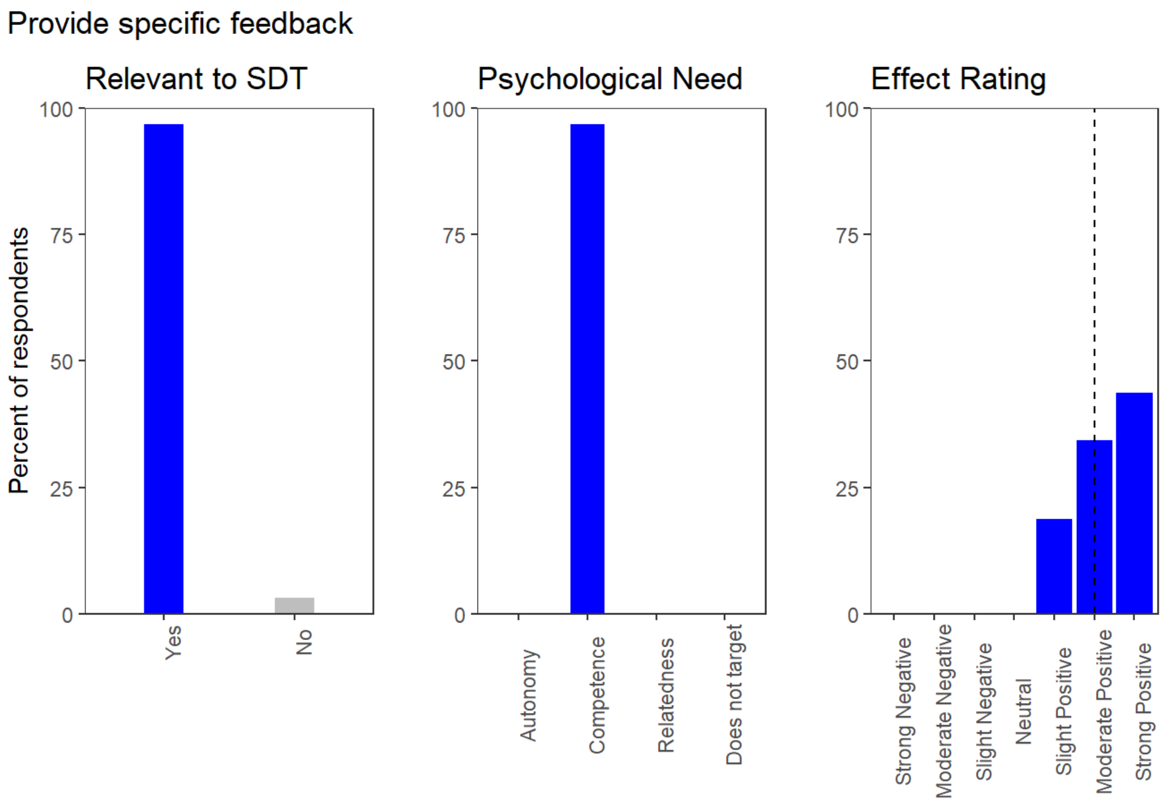
Provide feedback that targets a specific strategy for improvement

Example Behaviour:

"If you keep your eye on your attacker then you can try for an intercept, but mostly focus on marking your girl"

Function Description:

Clarifies path toward goal achievement.



TMB#10

Provide frequent feedback

Description:

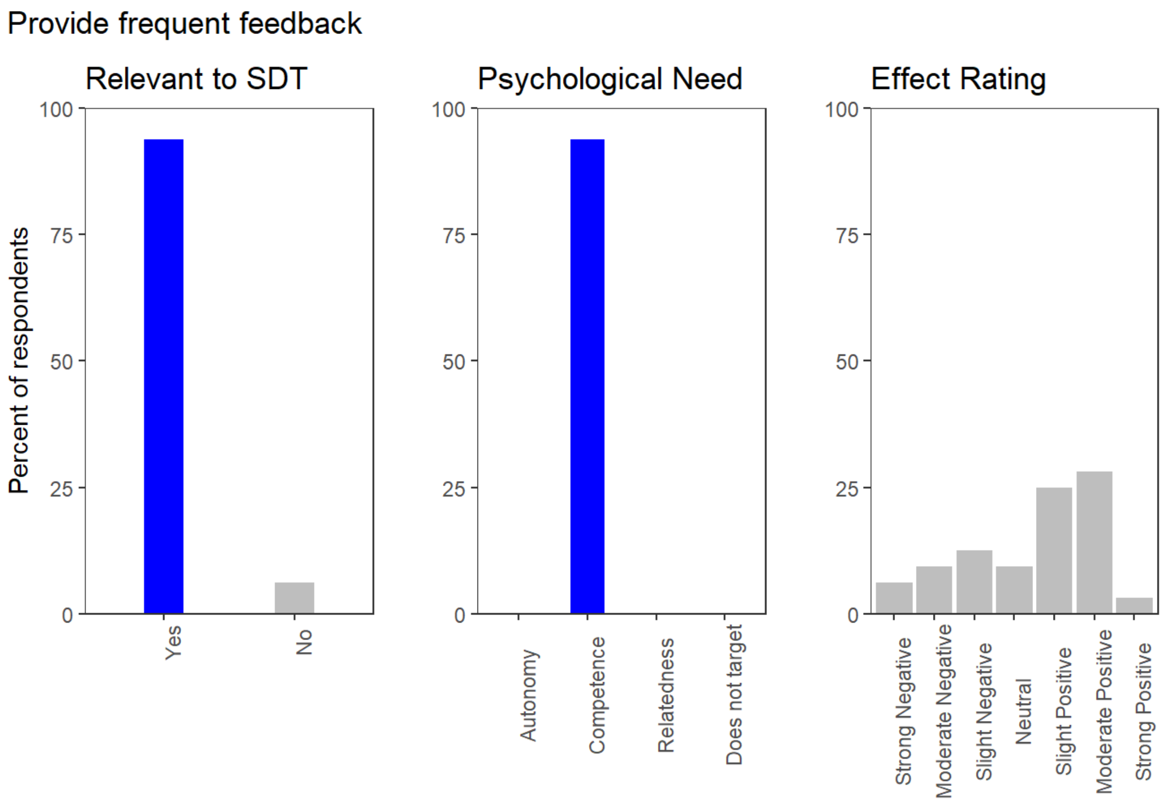
Frequently provide feedback to students

Example Behaviour:

Consistently monitoring the class to provide feedback for getting unstruck

Function Description:

Promotes continual improvement in abilities.



TMB#11

Offering hints

Description:

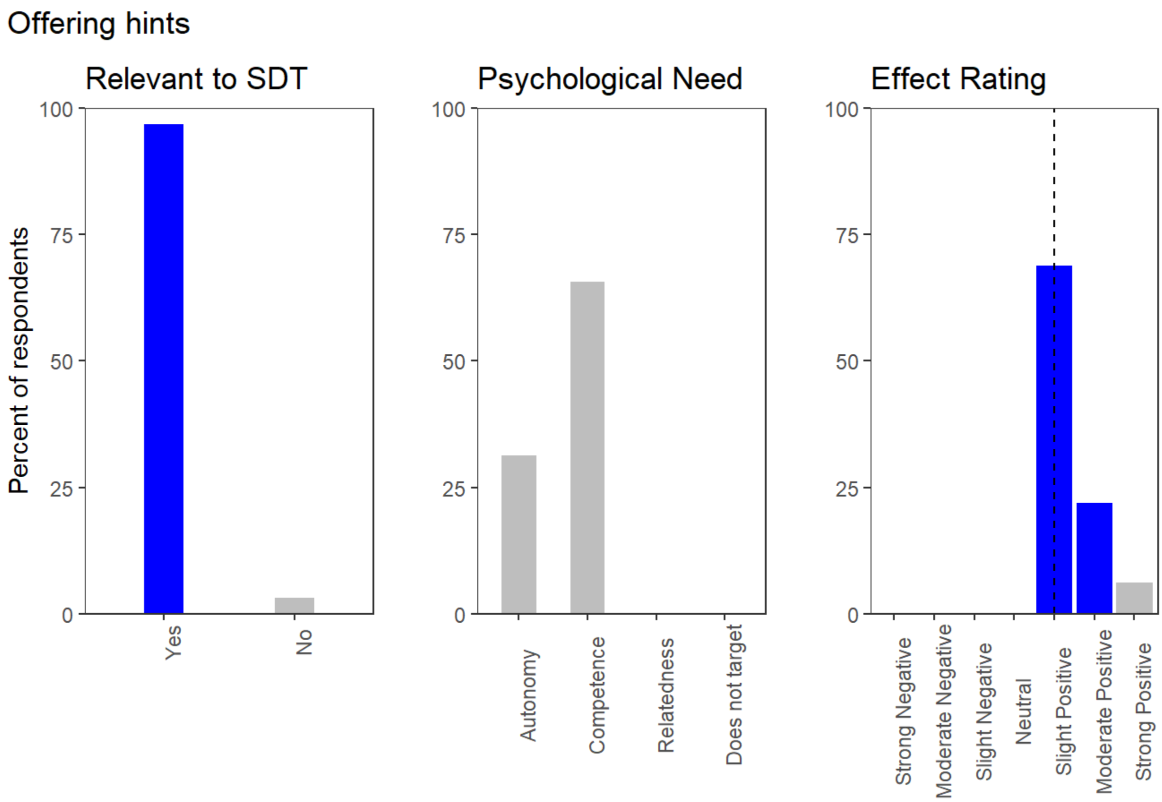
Give hints to help students along without giving them the "right answer"

Example Behaviour:

“It might be easier to start with this formula.”

Function Description:

Supports the student’s own learning processes. Allows students to maintain an internal locus of causality during learning.



TMB#12

Praise a student's fixed qualities

Description:

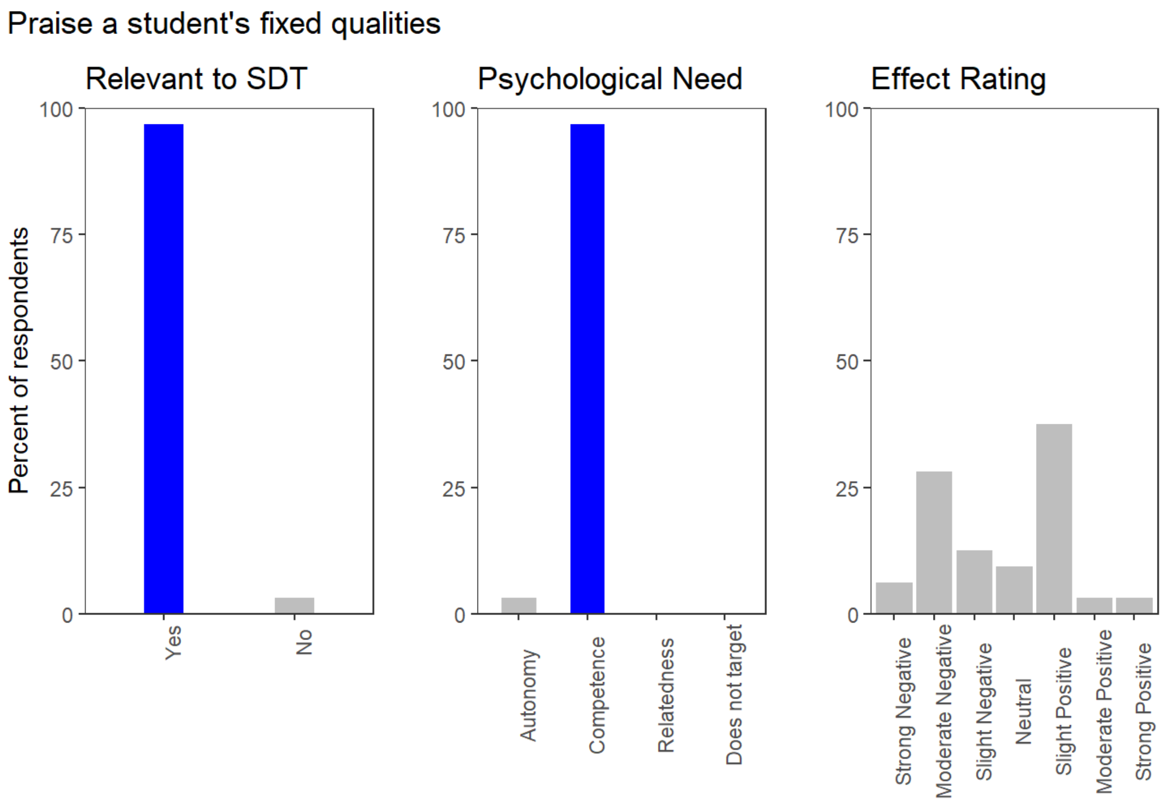
Provides praise that targets the talents or fixed qualities of students

Example Behaviour:

"You are very smart"

Function Description:

Affirms students natural abilities



TMB#13

Provide praise in public

Description:

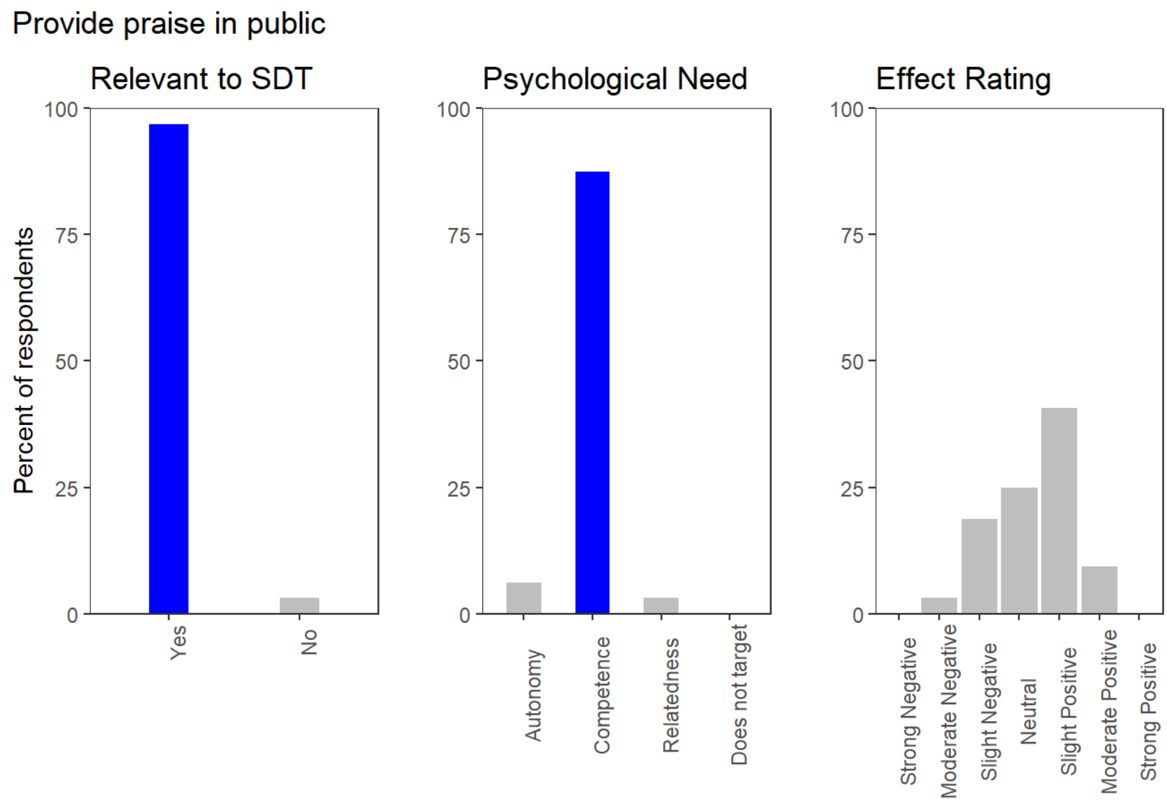
Praise a student in public

Example Behaviour:

Praise in front of the class

Function Description:

Generates pride within students receiving praise



TMB#14

Provide frequent praise

Description:

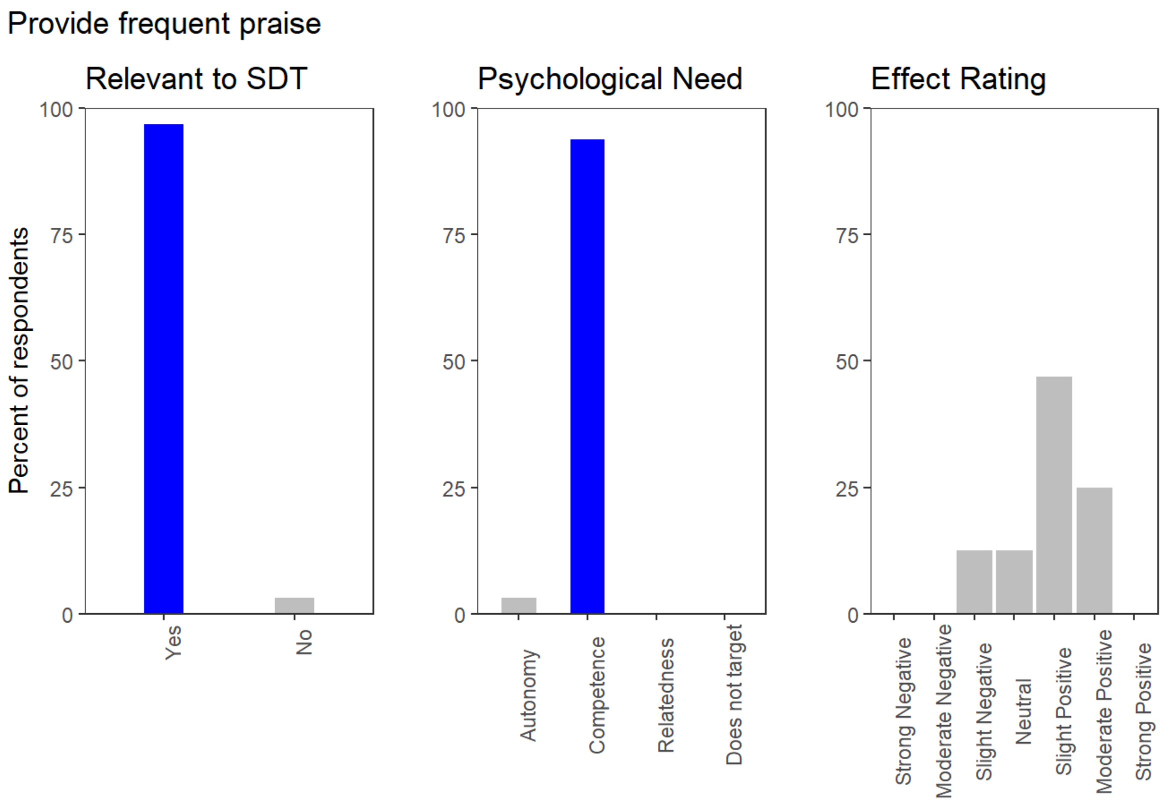
Frequently praise students

Example Behaviour:

Teacher consistently monitors the class, praising students for correct answers

Function Description:

Provides continual affirmation of progress and improvement



TMB#15

Provide extra resources for independent learning

Description:

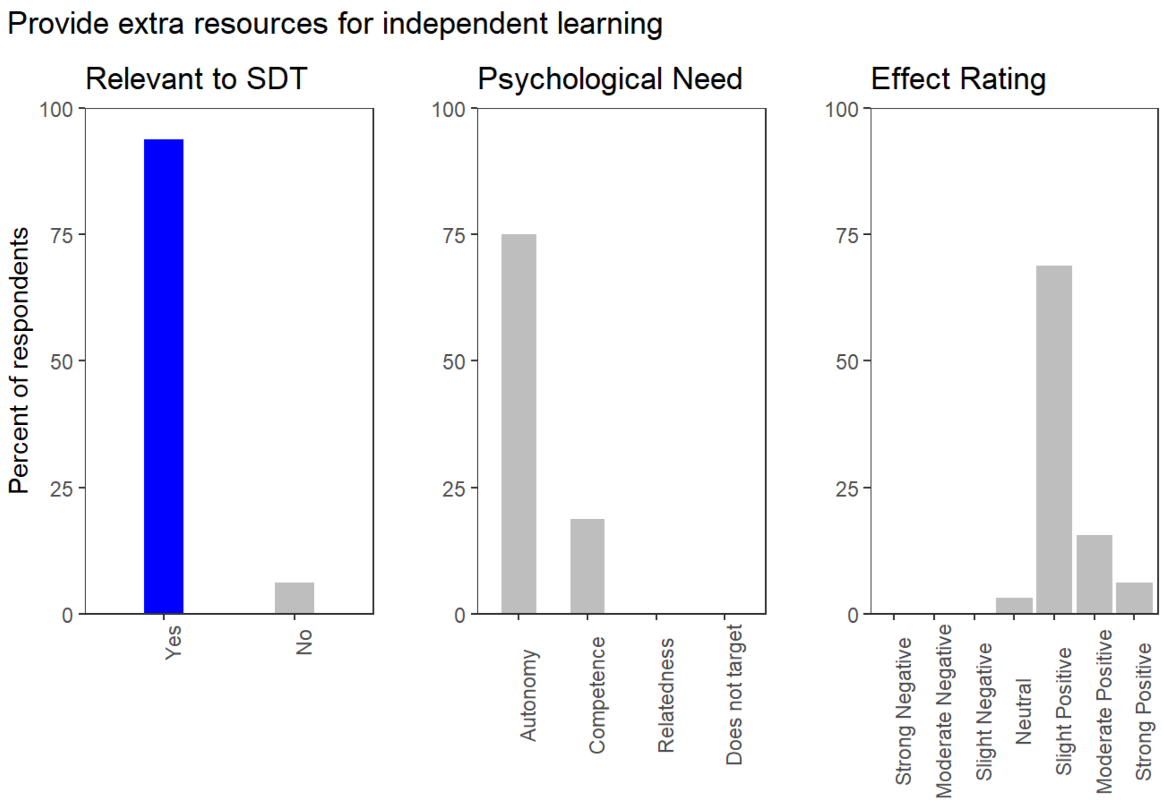
Introduce extra resources for further learning or support outside of class time

Example Behaviour:

"If you want more help, remember maths club before school tomorrow"; "here are some extra problems if you want to practice at home"

Function Description:

Allows for self-directed learning and progress outside of class time



TMB#16

Set goals based on self-refrenced standards

Description:

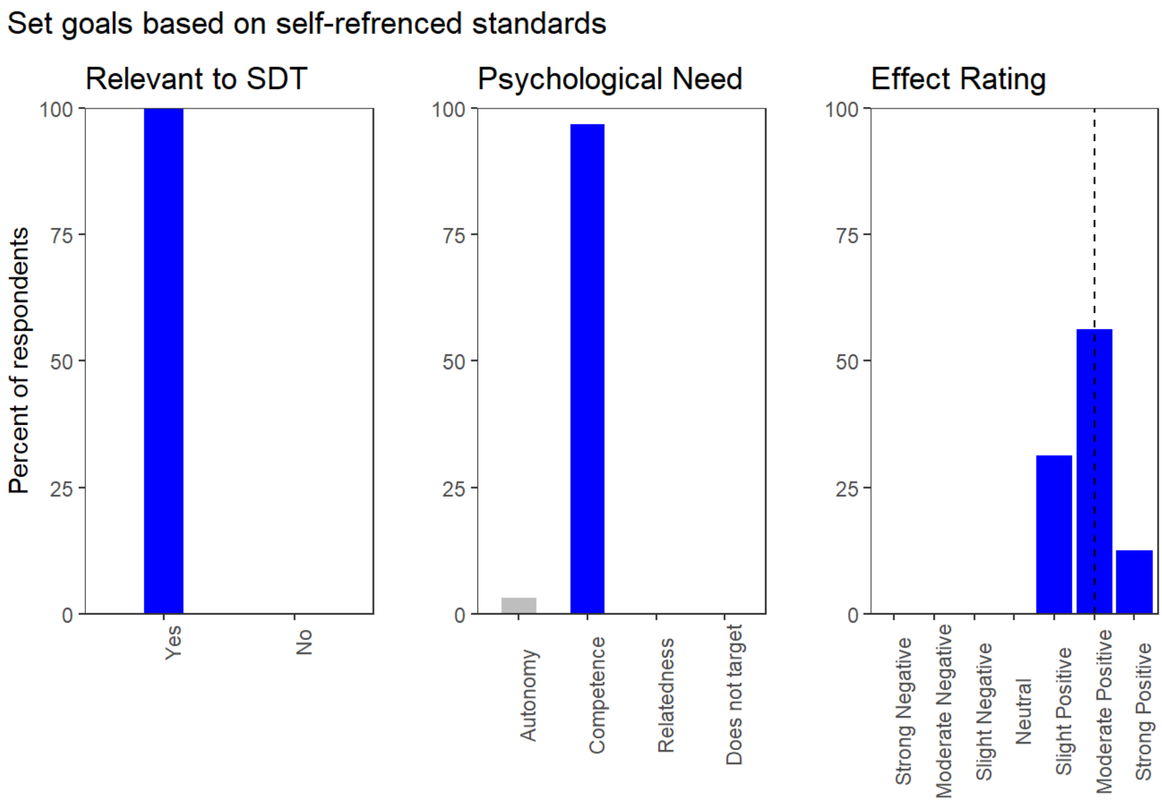
Set up activities where each student has their own goal

Example Behaviour:

"Try to jump futher than last time"

Function Description:

Promotes achievable goals by calibrating them to students skill



TMB#17

Clarify expectations

Description:

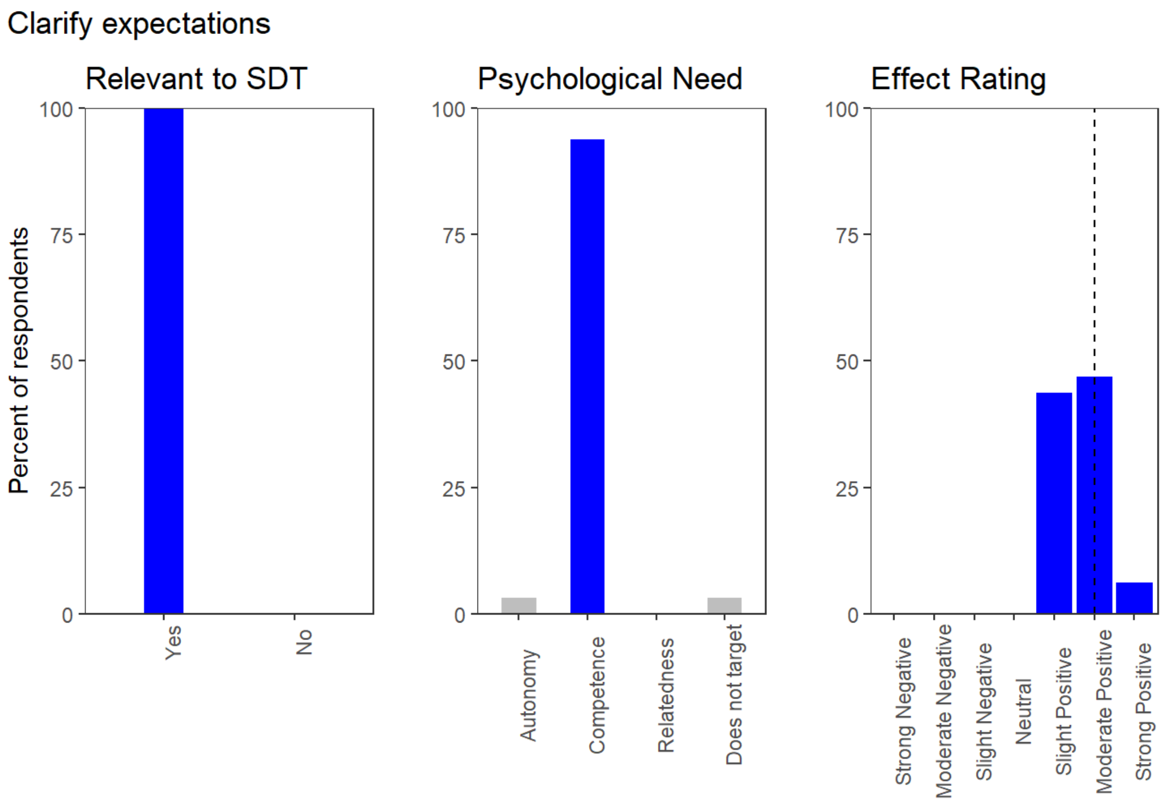
Provide clear instructions

Example Behaviour:

"Start with problems 4.1 to 4.4 then check your answers with me"

Function Description:

Provides structure so students know exactly what to do



TMB#18

Provide transparent structure

Description:

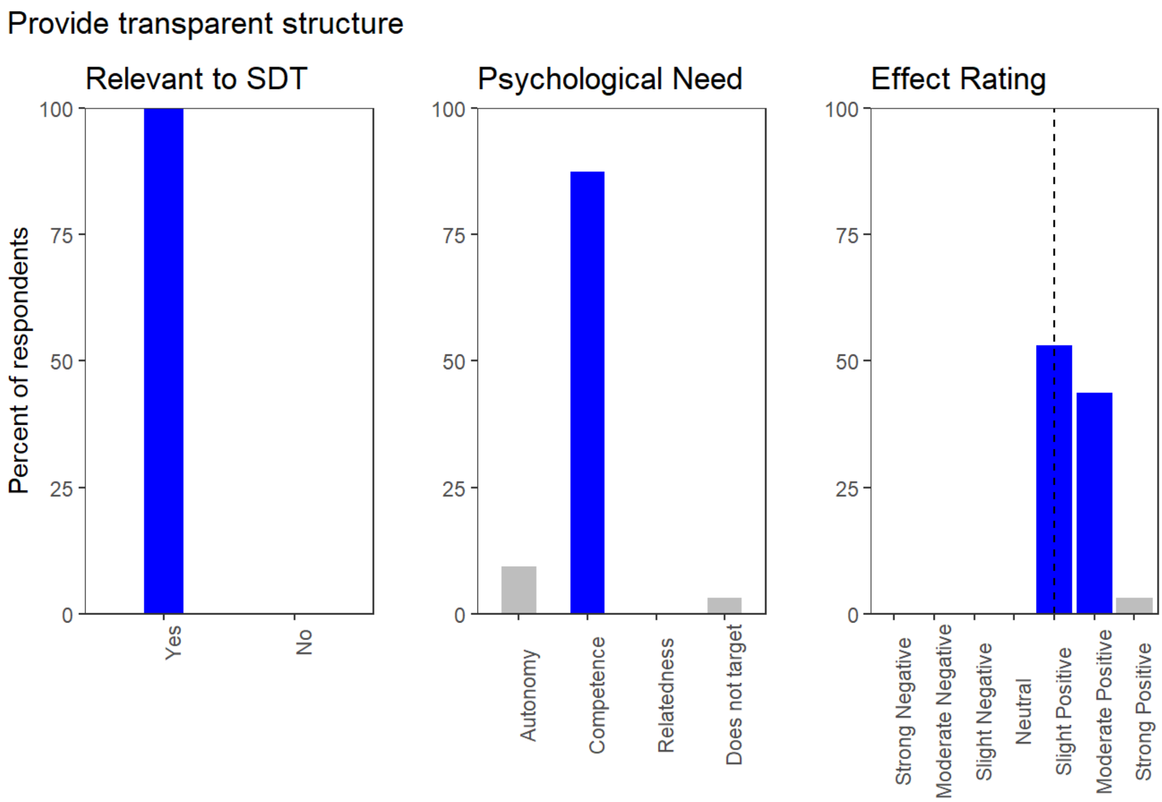
Provide an overview of what we are going to do in the lesson

Example Behaviour:

"In todays class, we are working on ratios in three ways..."

Function Description:

Provides a plan for students to follow so they know how things are going to be organised



TMB#19

Responding to Queries

Description:

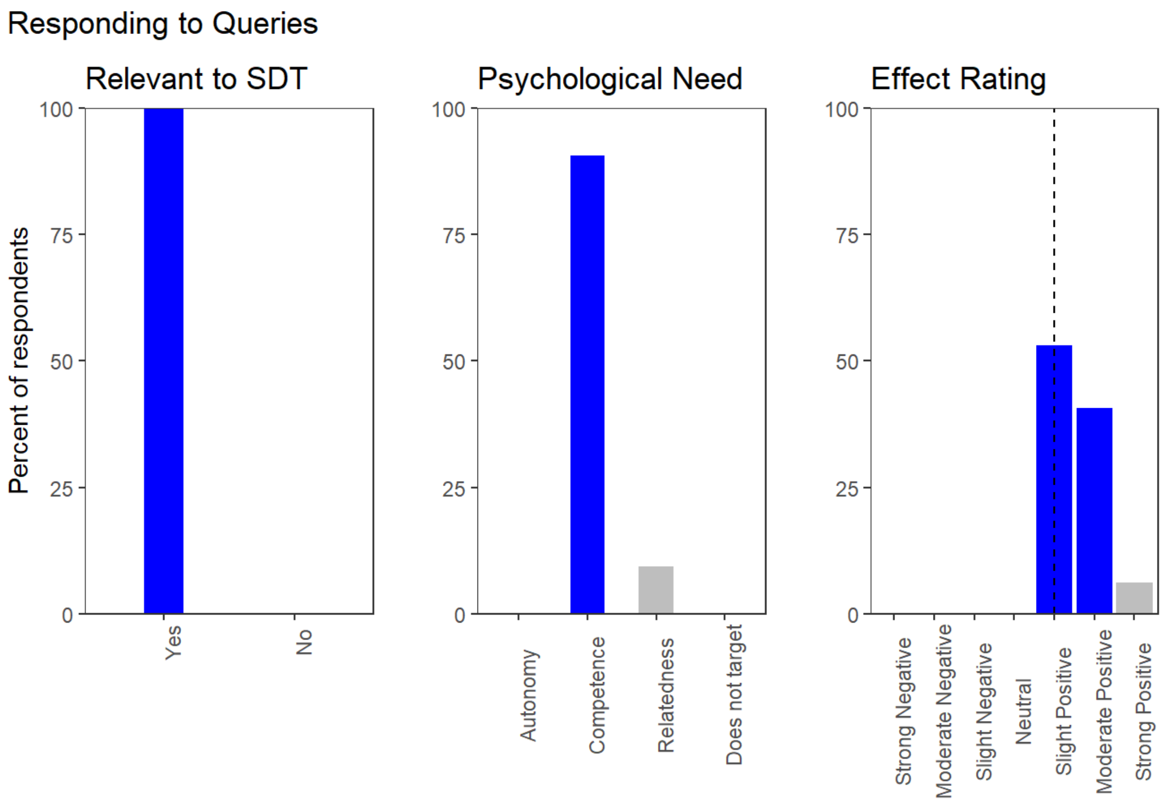
Answer student questions fully and carefully

Example Behaviour:

"No quite, that is the formula for Sin not Cos."

Function Description:

Clarifies path toward goal achievement.



TMB#20

Communicate in a perspective-taking way

Description:

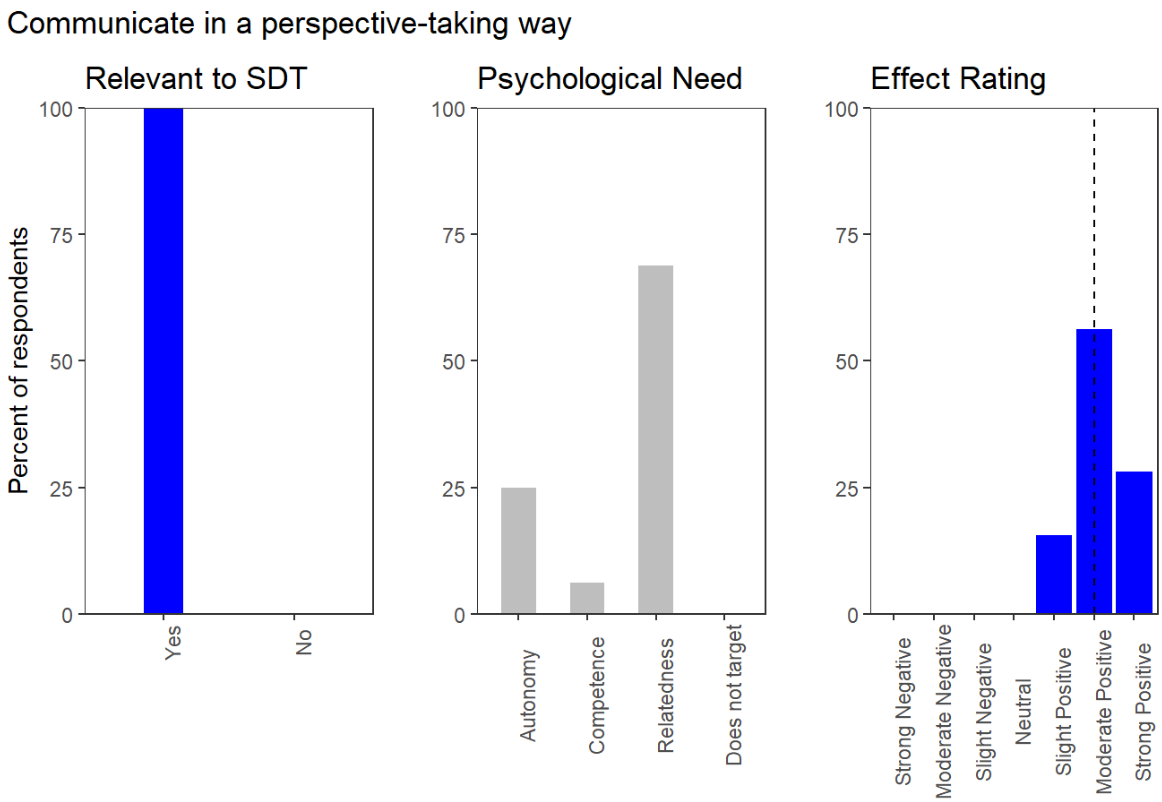
Show that you have taken a students perspective

Example Behaviour:

"Yes, you are right; this one is difficult"

Function Description:

Communicates that teacher understands the students frame of reference



TMB#21

Acknowledge student negative feelings

Description:

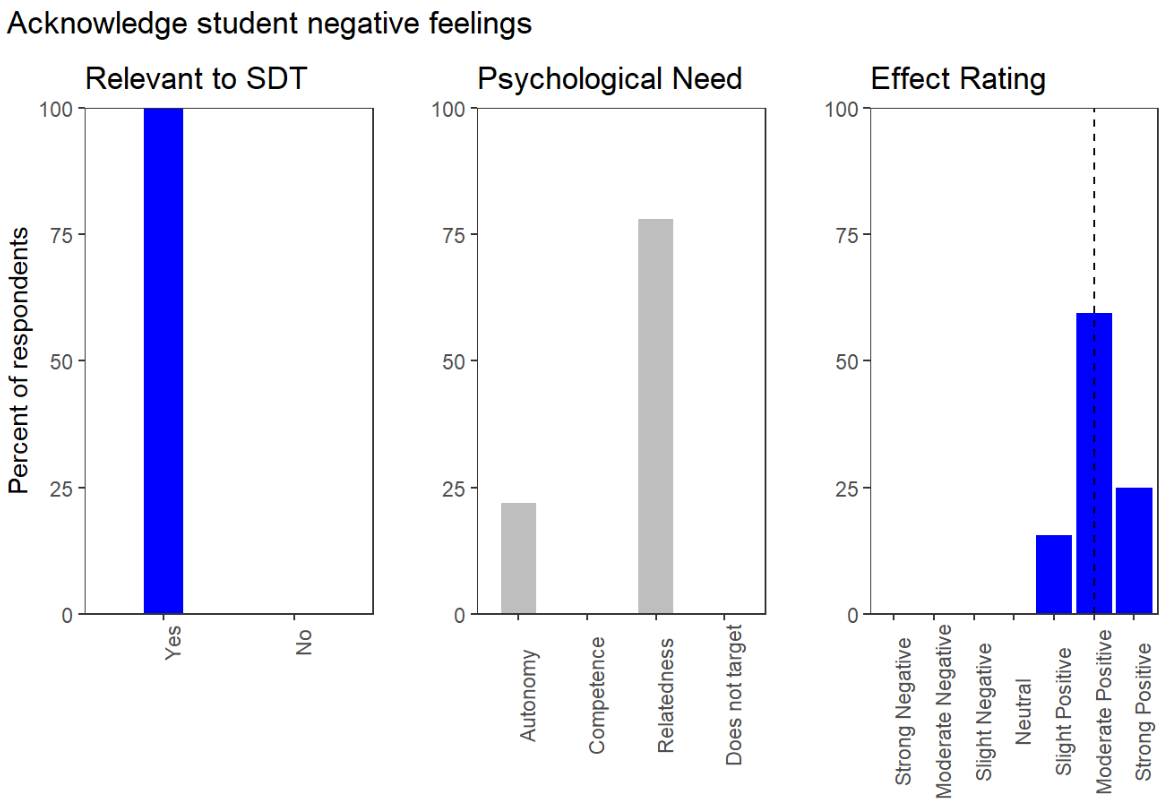
Acknowledge students negative feelings

Example Behaviour:

"I noticed you are looking frustrated."

Function Description:

Validates emotions as understandable, normal, and expected



TMB#22

Allow student own-paced progress

Description:

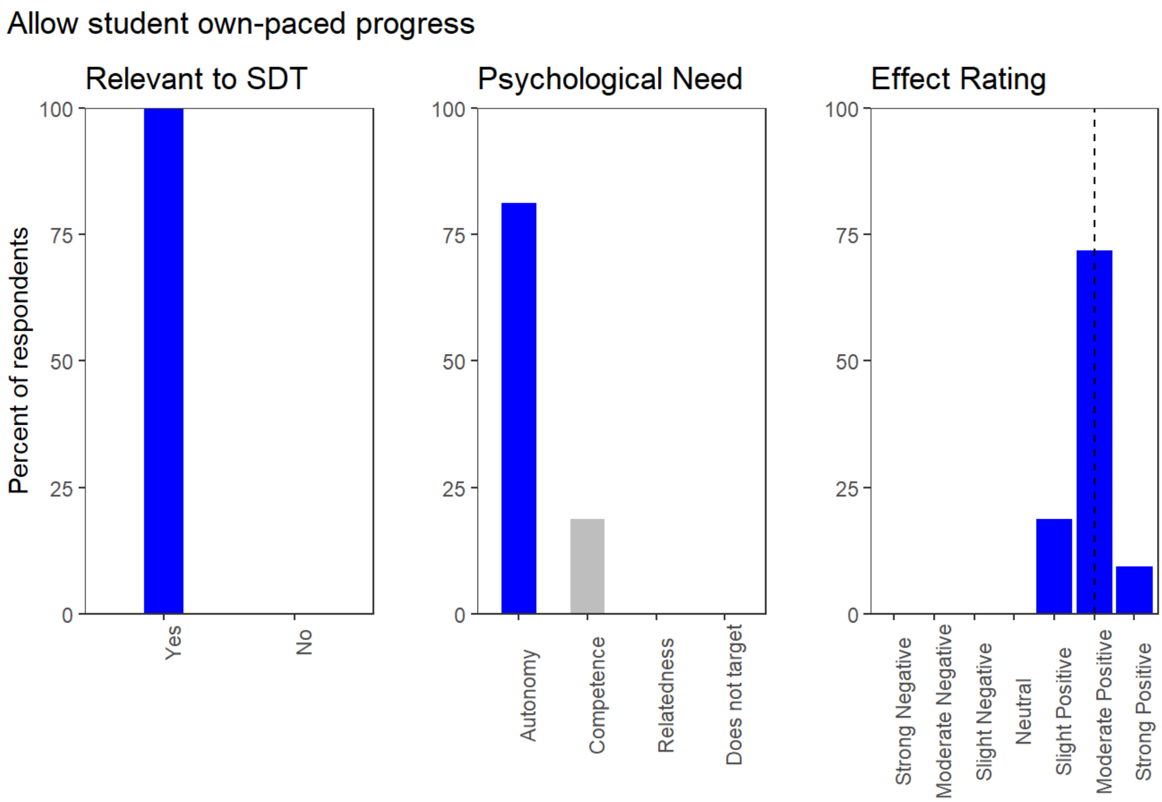
Allow the student to work independently and to solve a problem in his or her own pace

Example Behaviour:

"Solve the puzzle at your own pace"

Function Description:

Lets students manage their own cognitive load so they do not get frustrated or overwhelmed



TMB#23

Teach in students' preferred ways

Description:

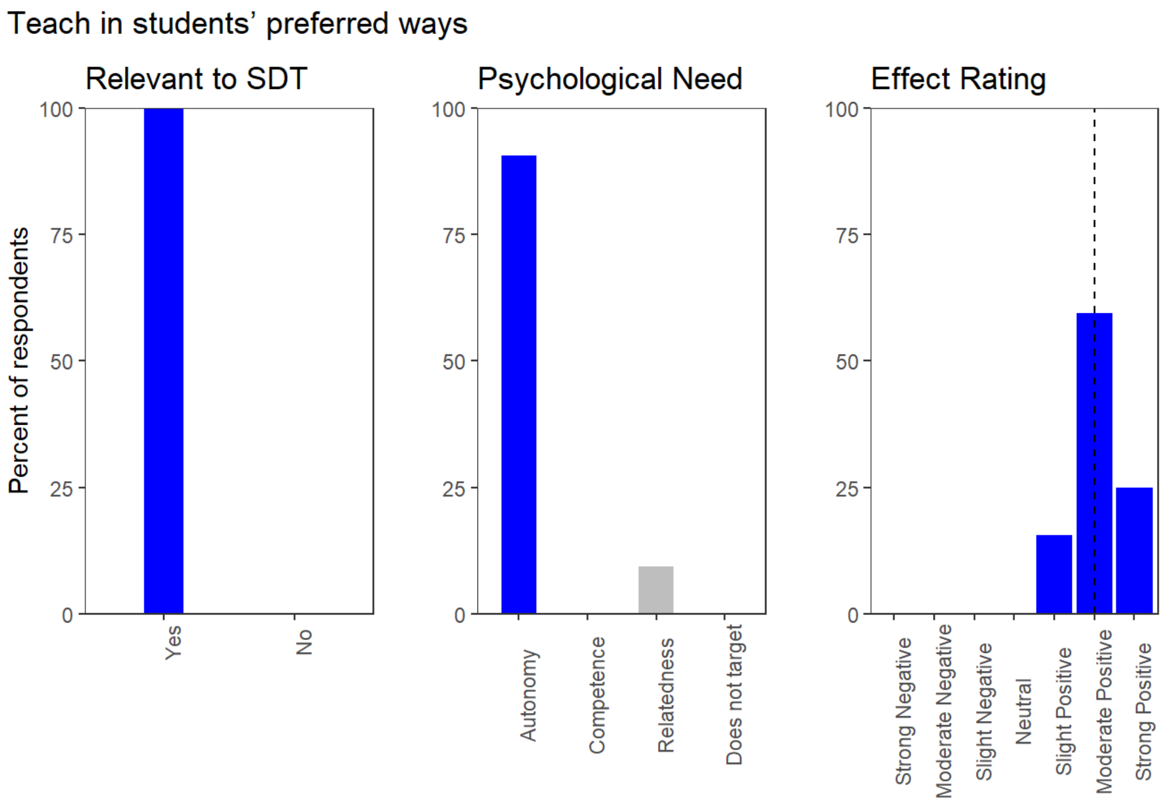
Use knowledge gleaned about the student values and preferences to design class activities customised to them.

Example Behaviour:

"I know you love comics so I based todays lesson on ..."

Function Description:

Aligns lesson activities to students intrinsic reasons for learning rather than imposing extrinsic reasons



TMB#24

Provoke curiosity

Description:

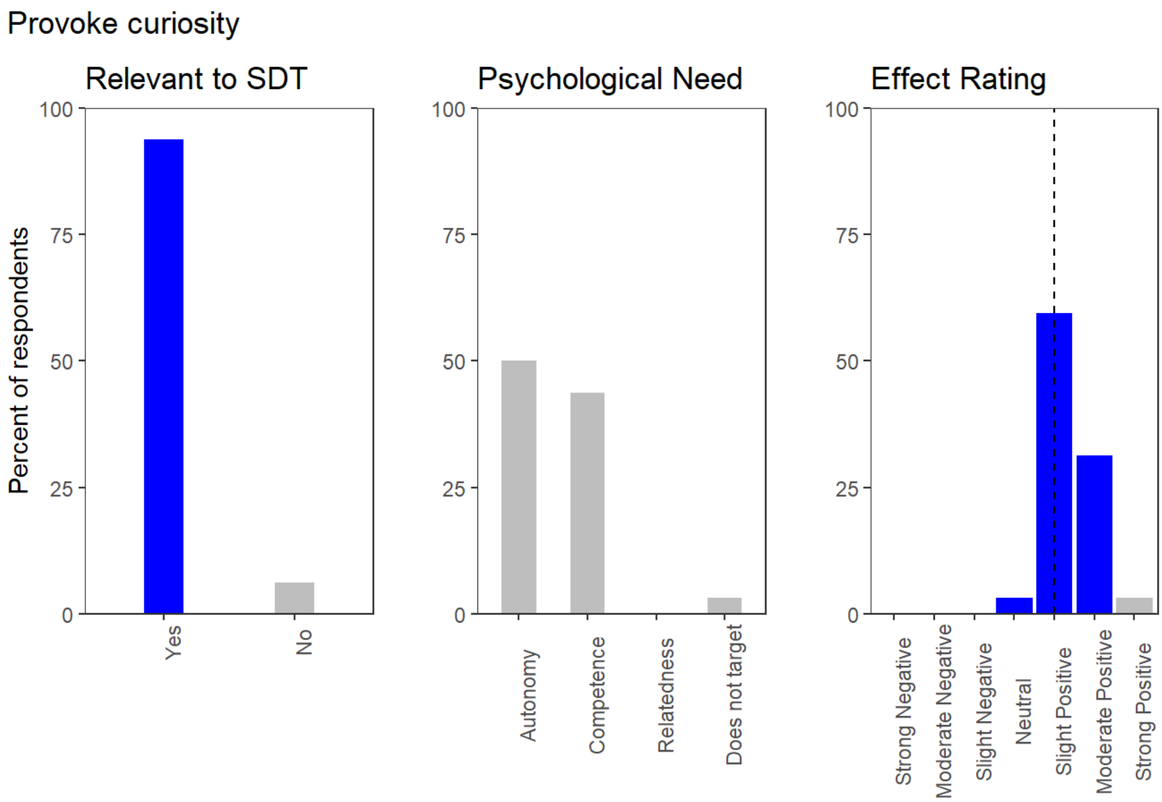
Ask a curiosity-inducing question

Example Behaviour:

"How long does it take the Earth to go around the sun?"

Function Description:

Supports students competence through facilitating their exploratory behaviour



TMB#25

Display explicit guidance

Description:

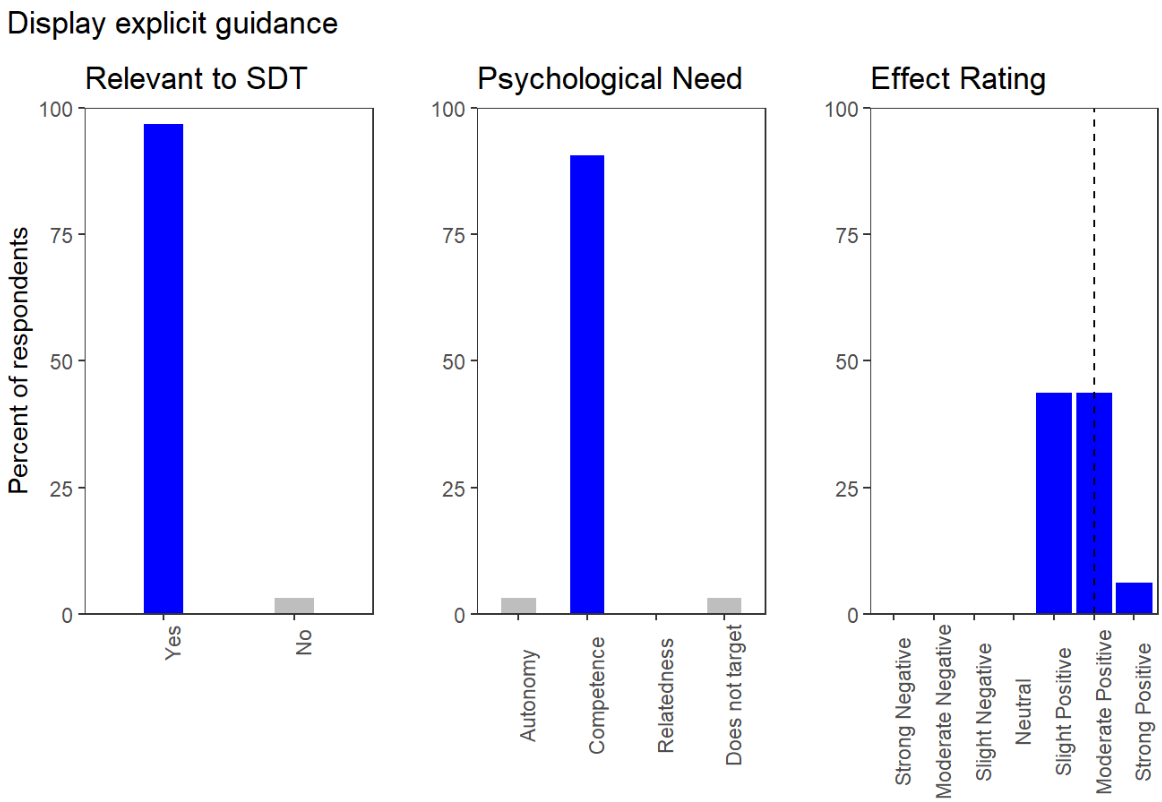
Provide clear guidance, clear goal, and clear action plans

Example Behaviour:

"To understand how volcanoes work, we're going to make a model. First, grab a test-tube, some vinegar, and some baking soda."

Function Description:

Enables students to understand success criteria



TMB#26

Show understanding of the students' point of view

Description:

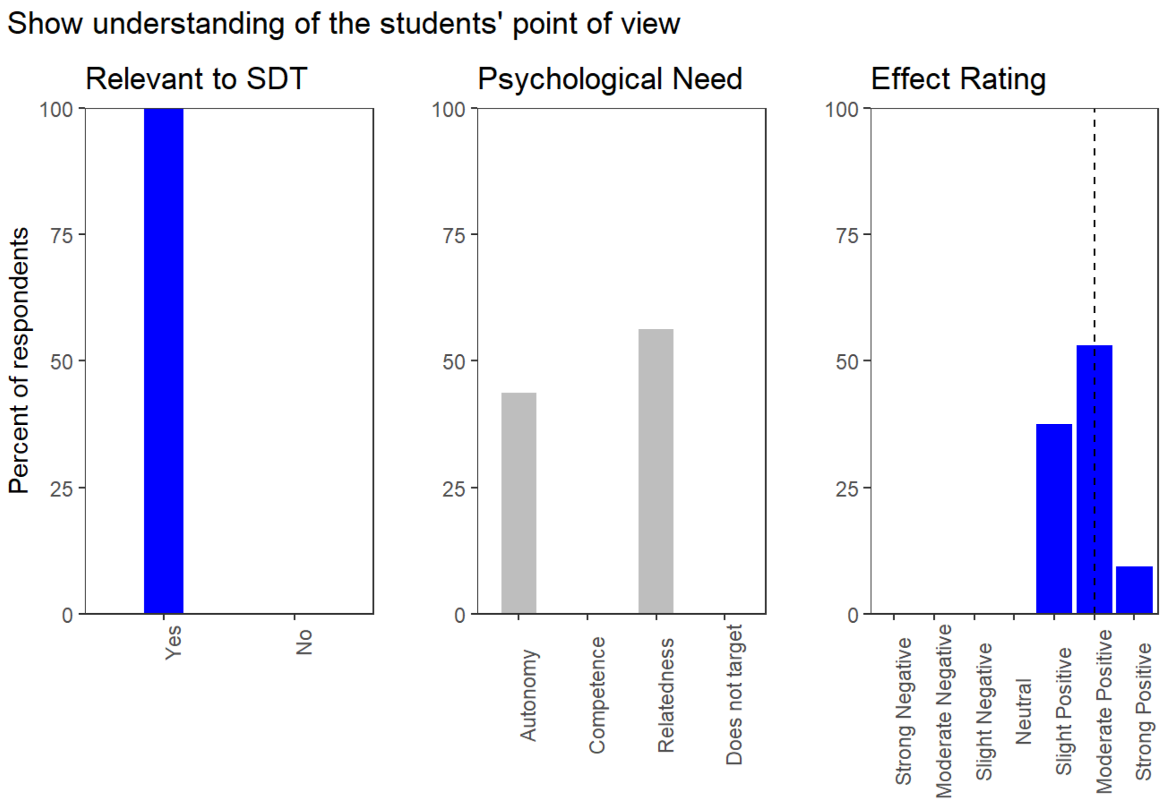
Try to understand how students see things before suggesting a new way to do things.

Example Behaviour:

"You probably hate fitness tests and can not see the point."

Function Description:

Helps the student feel listened-to and understood.



TMB#27

Use pupils as positive role models

Description:

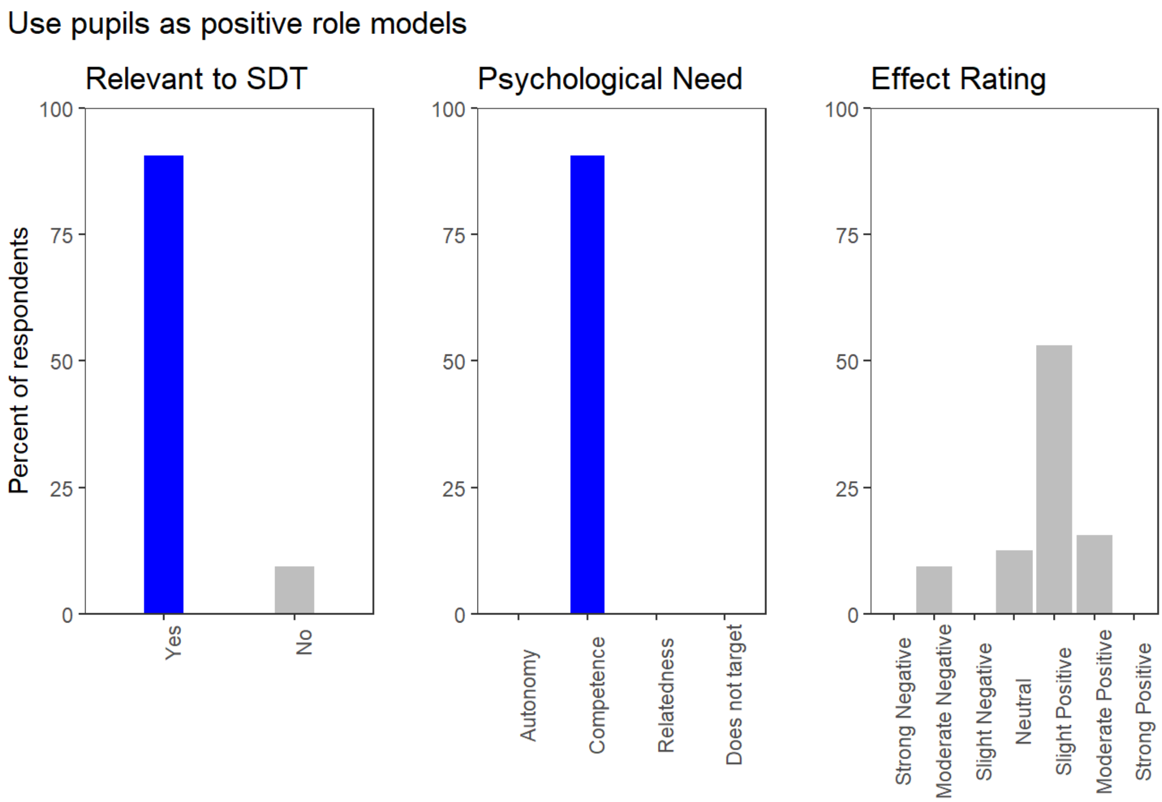
Highlight some students as examples for the rest of the class to follow

Example Behaviour:

"John, you commented on your code very well. Can we put it on the smartboard so your friends can see it?"

Function Description:

Increase self-belief through vicarious experiences of success



TMB#28

Self-monitoring of progress and effort

Description:

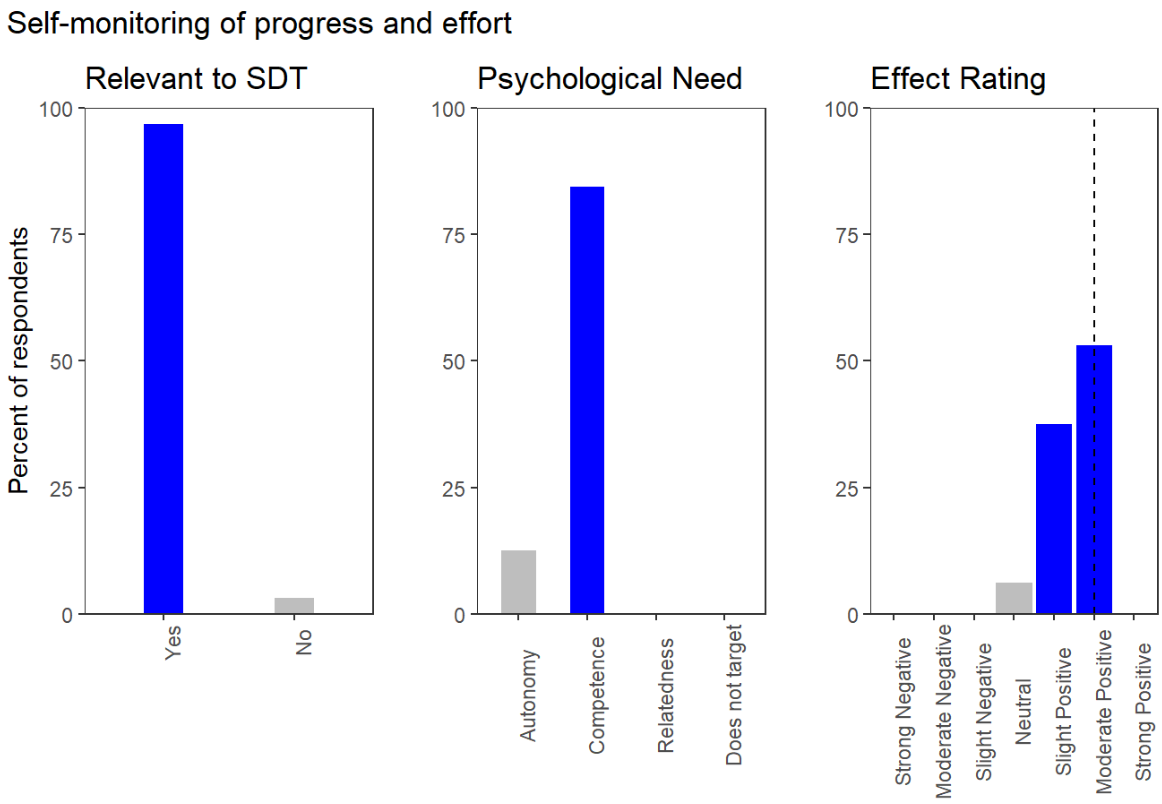
Facilitate monitoring of progress, skill level, or performance

Example Behaviour:

"How would you rate your performance in the last three weeks?"

Function Description:

Provides opportunities for accurate self-reflection of effort and progress



TMB#29

Undifferentiated challenge

Description:

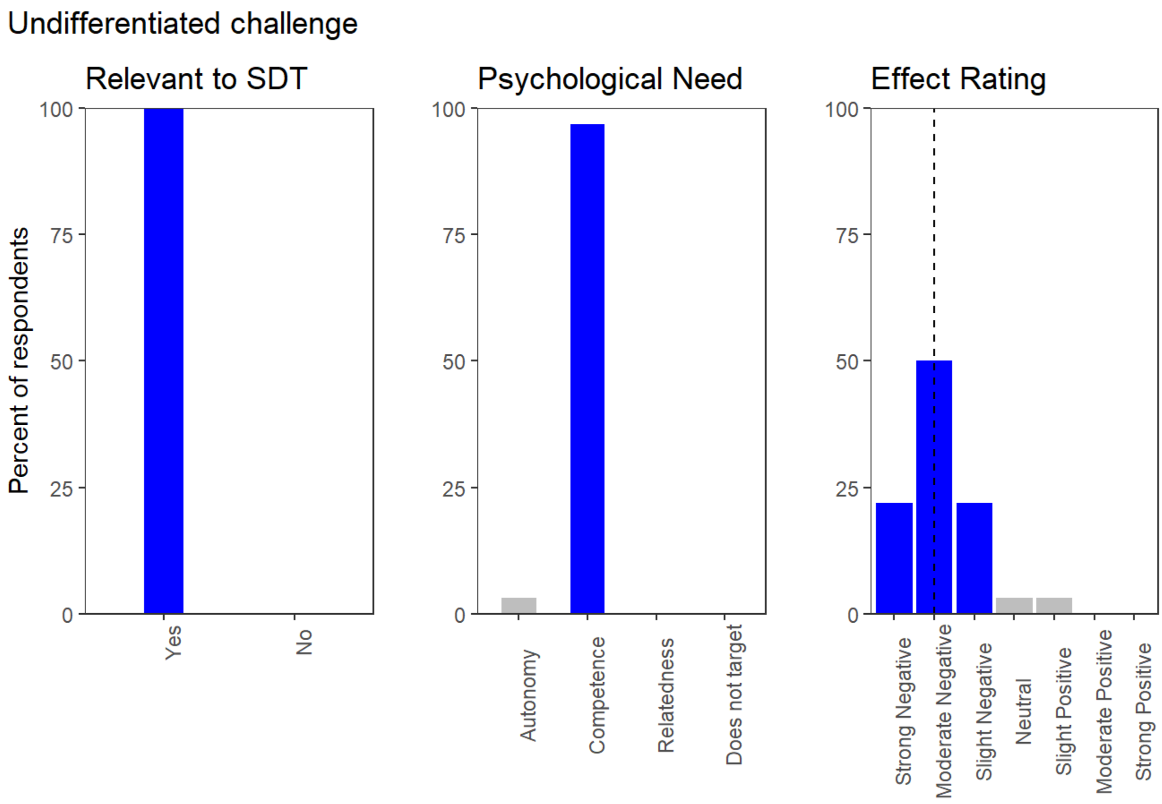
The same task is set for all students regardless of their level of ability.

Example Behaviour:

"Try to do a lay up by using the backboard"

Function Description:

Given natural variation in abilities, many students may be bored and others overwhelmed.



TMB#30

Ask controlling questions

Description:

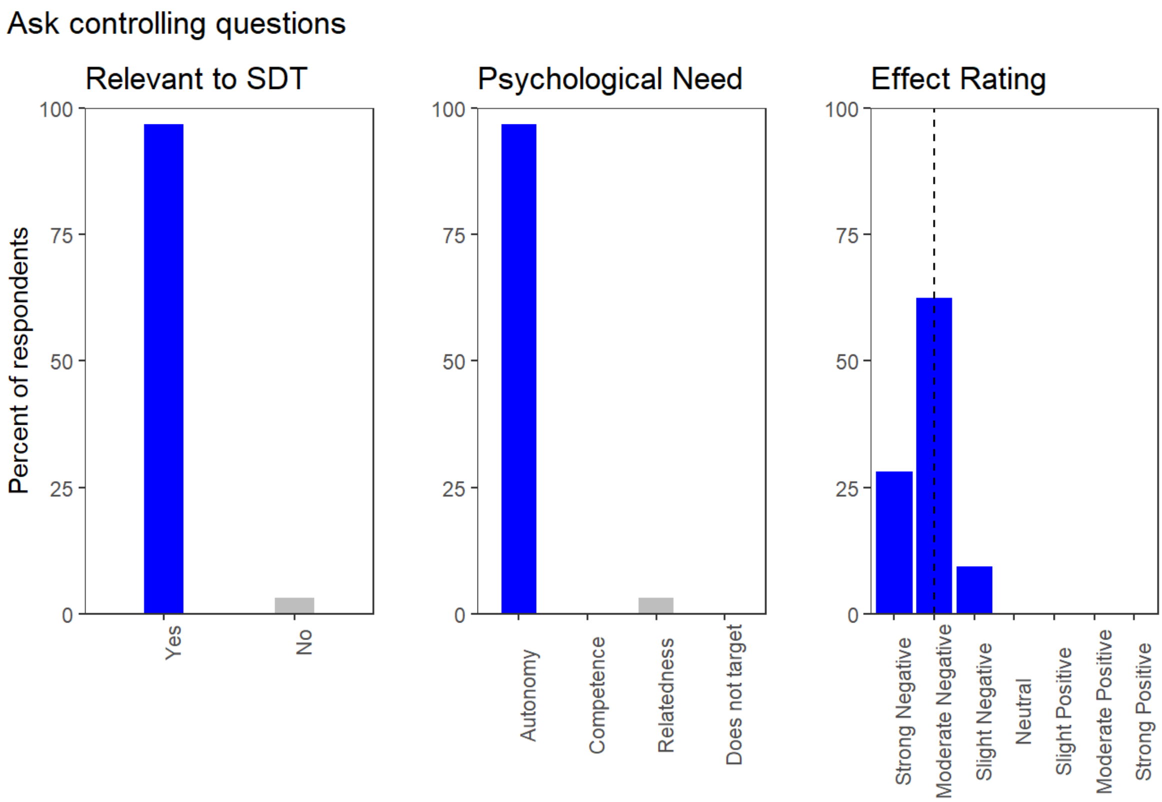
Provide commands that are phrased as rhetorical questions

Example Behaviour:

"Do it like I showed you!"

Function Description:

Communicates disapproval for the students current behaviour without clarifying how to improve or a rationale for change.



TMB#31

Set challenging deadlines

Description:

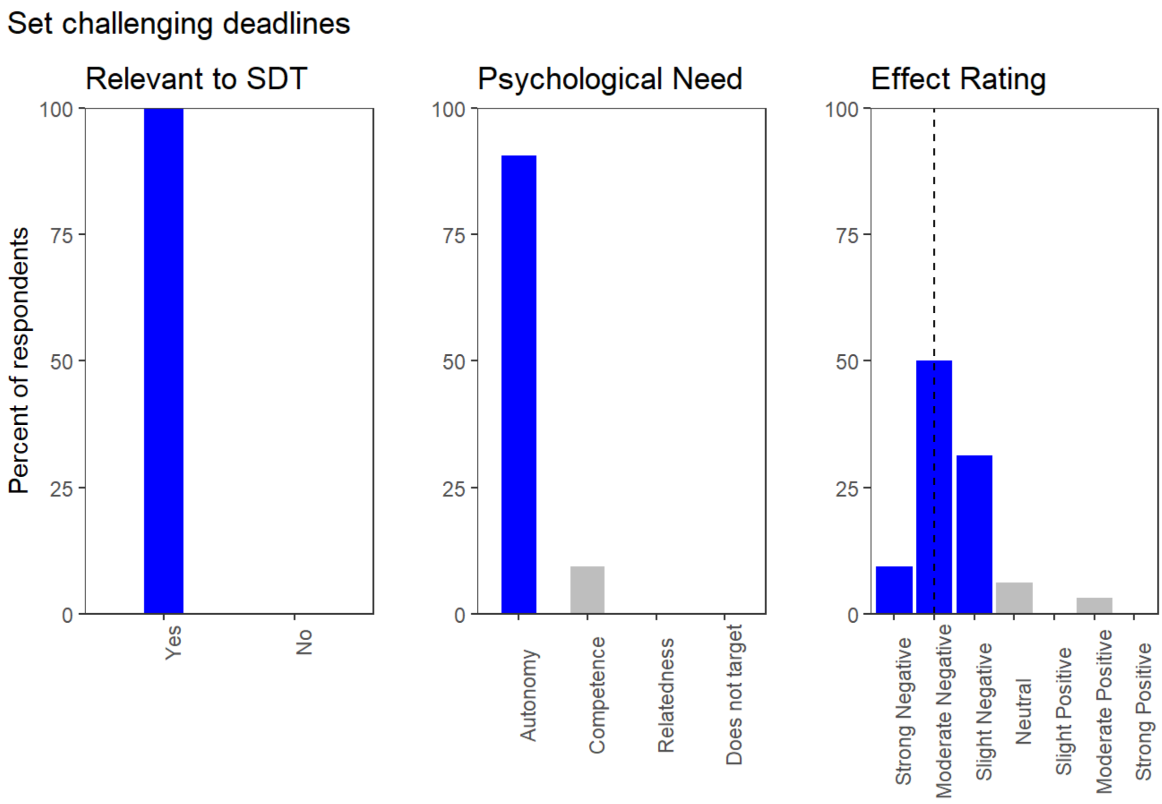
Allow a capped amount of time for a task, or remind students they are running out of time

Example Behaviour:

"Spend 10 minutes on this worksheet"; "We only have a few minutes left"

Function Description:

Adds pressure on students to work faster and finish tasks when the teachers says to



TMB#32

Outlining Punishment Contingencies

Description:

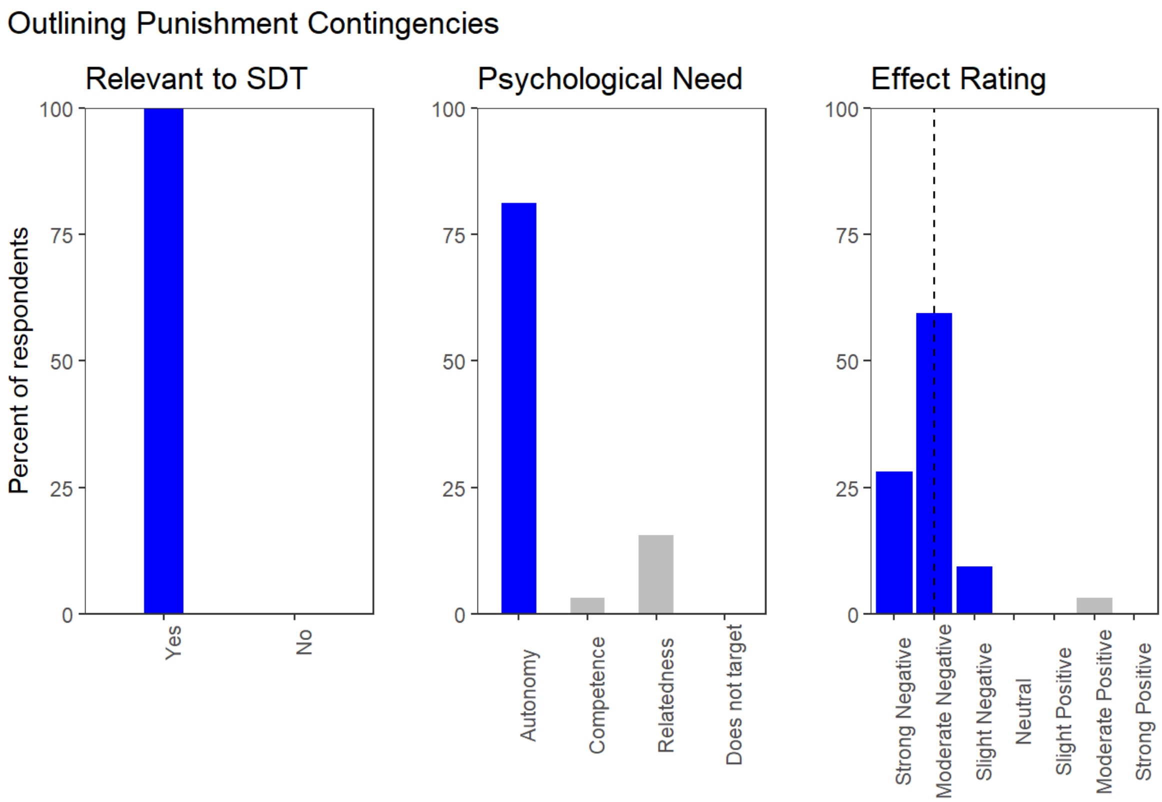
Declaring (but not yet enforcing) if-then extrinsic punishments—contingencies that are not inherent to the task and are provided in an effort to extinguish a behaviour

Example Behaviour:

"If you two speak one more time, I will send you out"

Function Description:

Imposes an extrinsic reason for student behaviour.



TMB#33

Provide rewards fairly

Description:

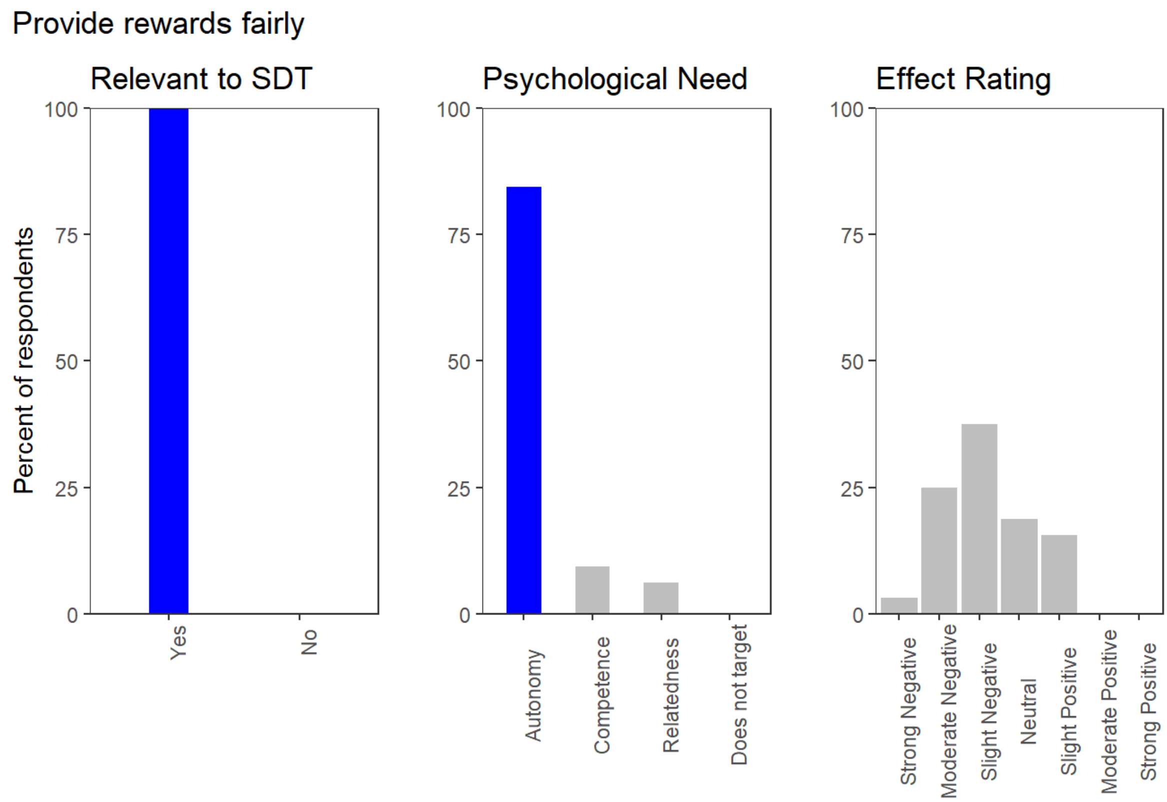
Provide rewards when the expected behaviour is observed

Example Behaviour:

"You all did your homework, so as I promised, we can watch a YouTube video today"

Function Description:

Adds external, tangible signal of which behaviours are desirable/valued by the teacher



TMB#34

Use abusive language (content)

Description:

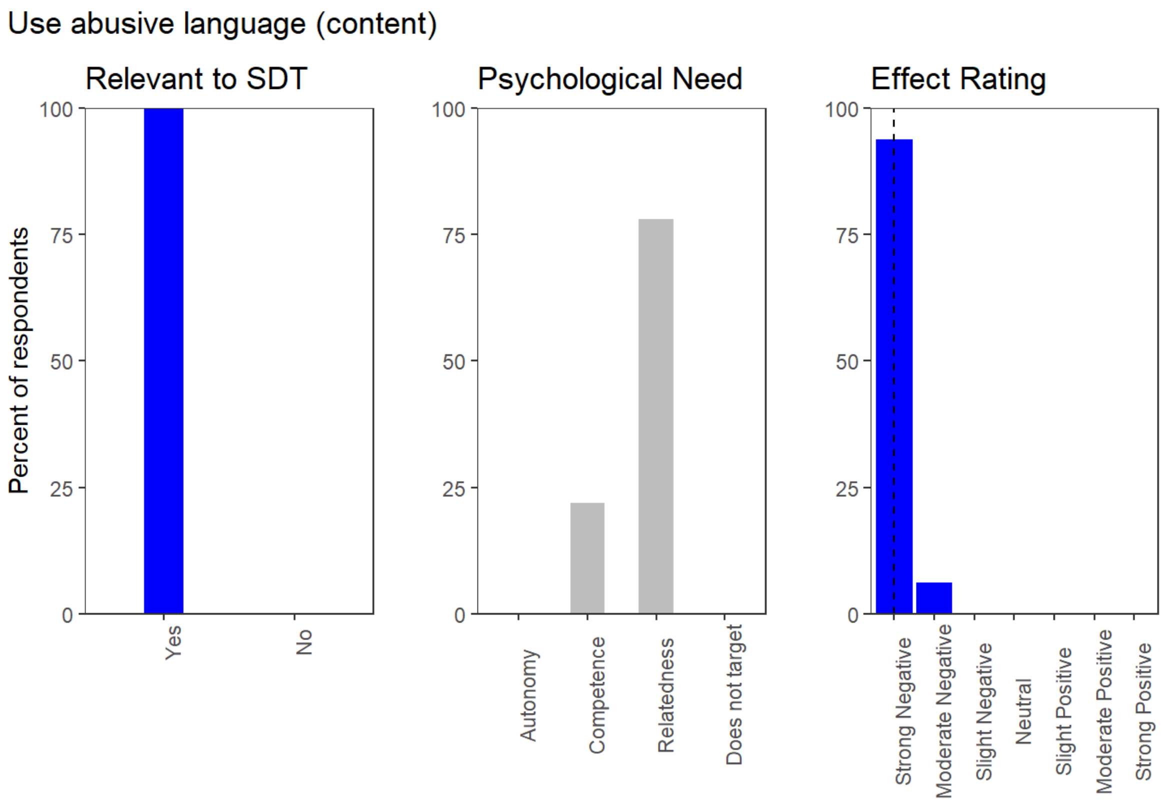
Calling students by hurtful names when they misbehave

Example Behaviour:

Calling a student "dummie" or "moron"

Function Description:

Performance mistakes and behavioural misconduct are met with competence-threatening punishment



TMB#35

Yell or use a harsh tone

Description:

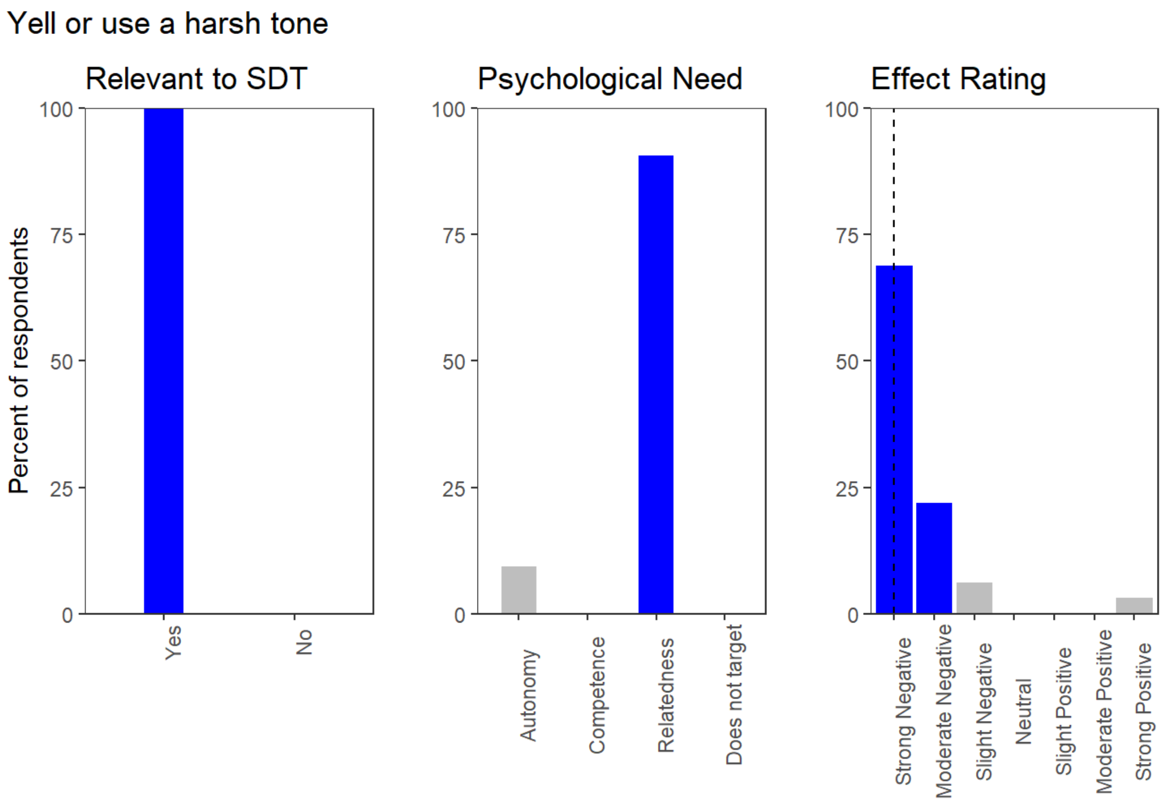
Teacher yells to get control of the classs

Example Behaviour:

Yelling such as "HEY!"; "STOP IT!"

Function Description:

Creates a more emotionally unstable and unpredictable environment for students, increasing fear



TMB#36

Provide rewards unfairly

Description:

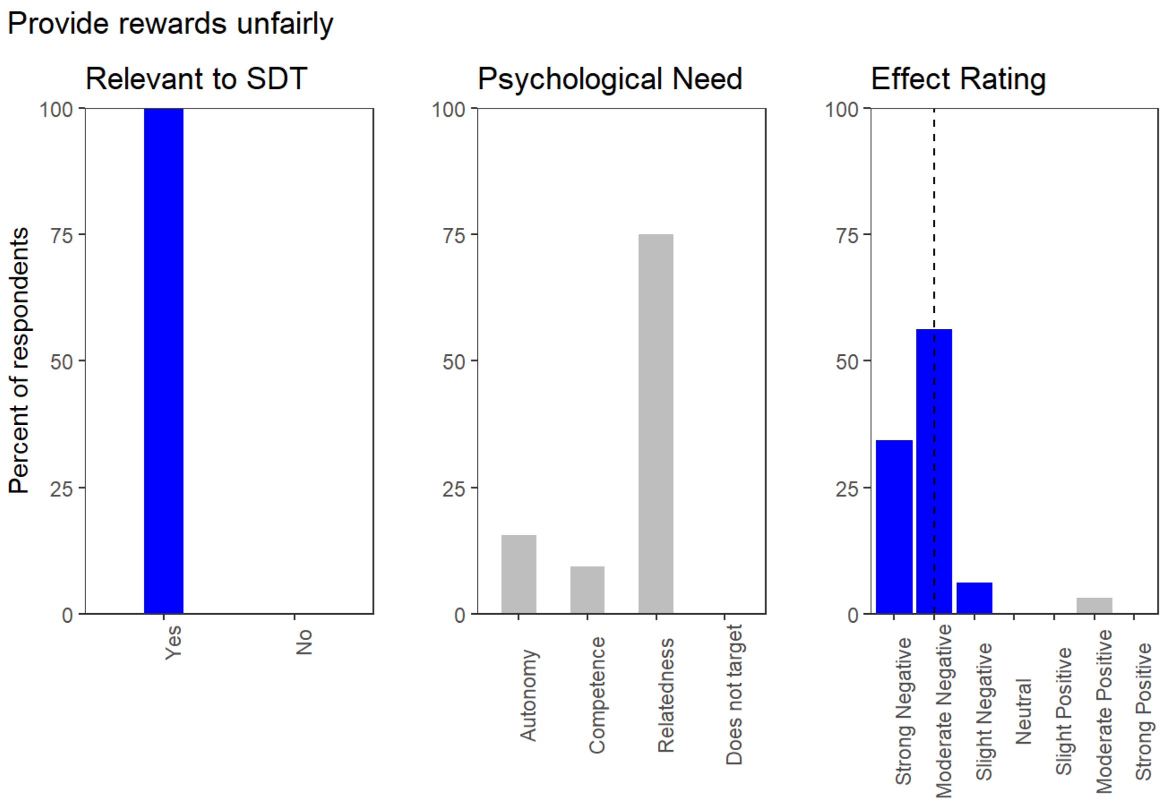
Provide rewards unfairly so students who are doing equally well, get different rewards

Example Behaviour:

Rewarding only one of three people who all completed a task

Function Description:

Students feel rewards are not predictable and teacher behaviour unjust



TMB#37

Provide punishments unfairly

Description:

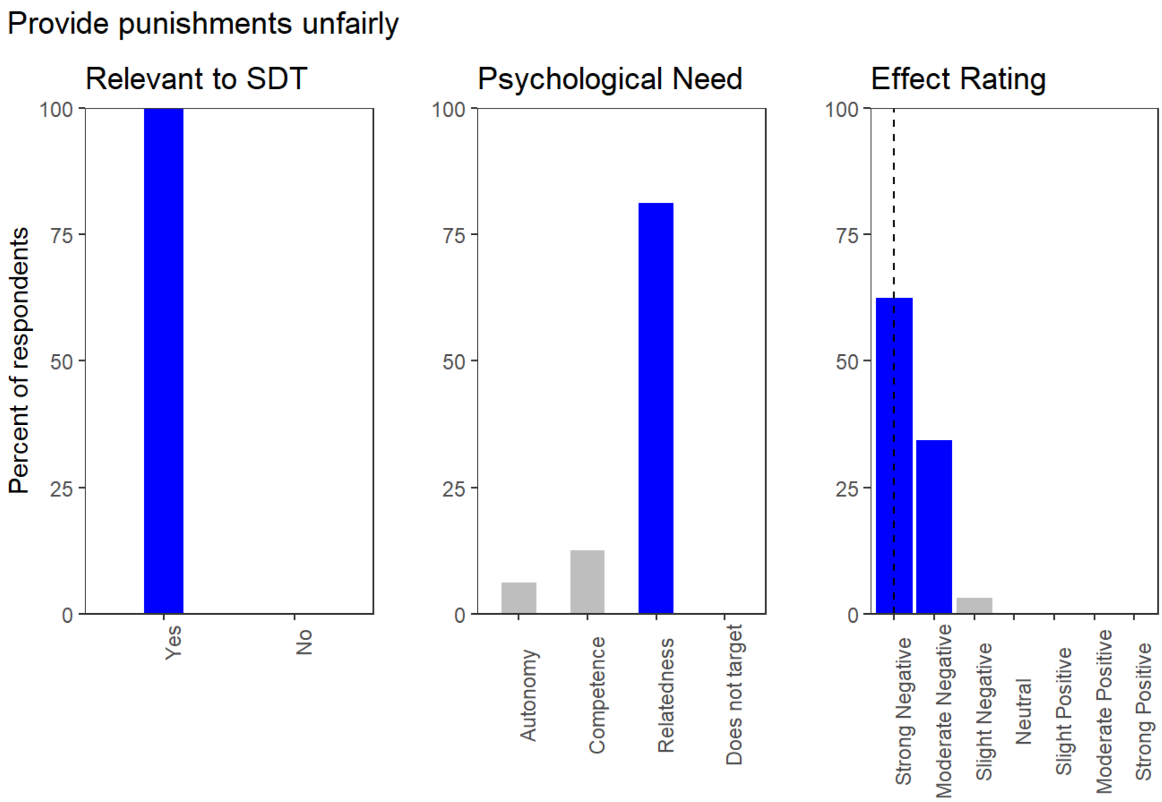
Provide punishments unfairly so students who misbehave are treated unequally

Example Behaviour:

Punishing only one of two students who are speaking out of turn

Function Description:

Means structures are perceived as inconsistent and unreliable



TMB#38

Publicly present critical feedback

Description:

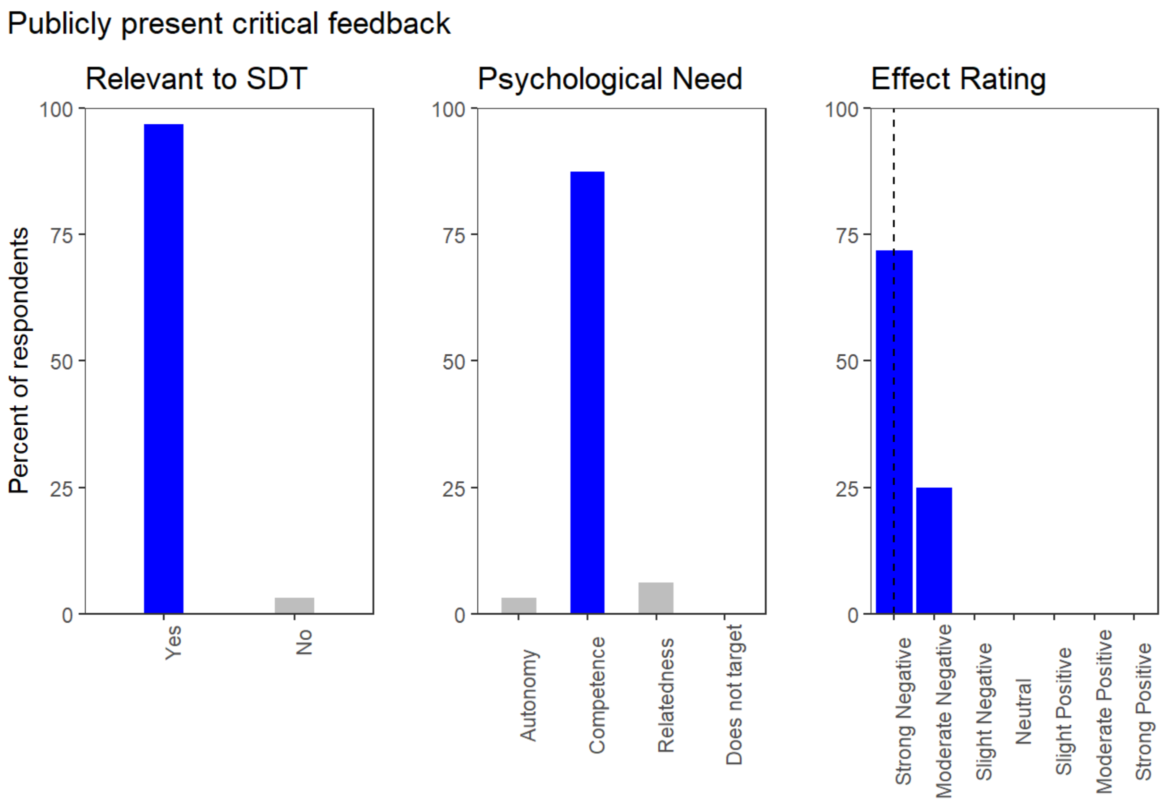
Provide critical feedback in public so other students can hear

Example Behaviour:

Provide critical feedback in front of the class

Function Description:

Increases risk of feedback being ego-threatening



TMB#39

Use vague criticism

Description:

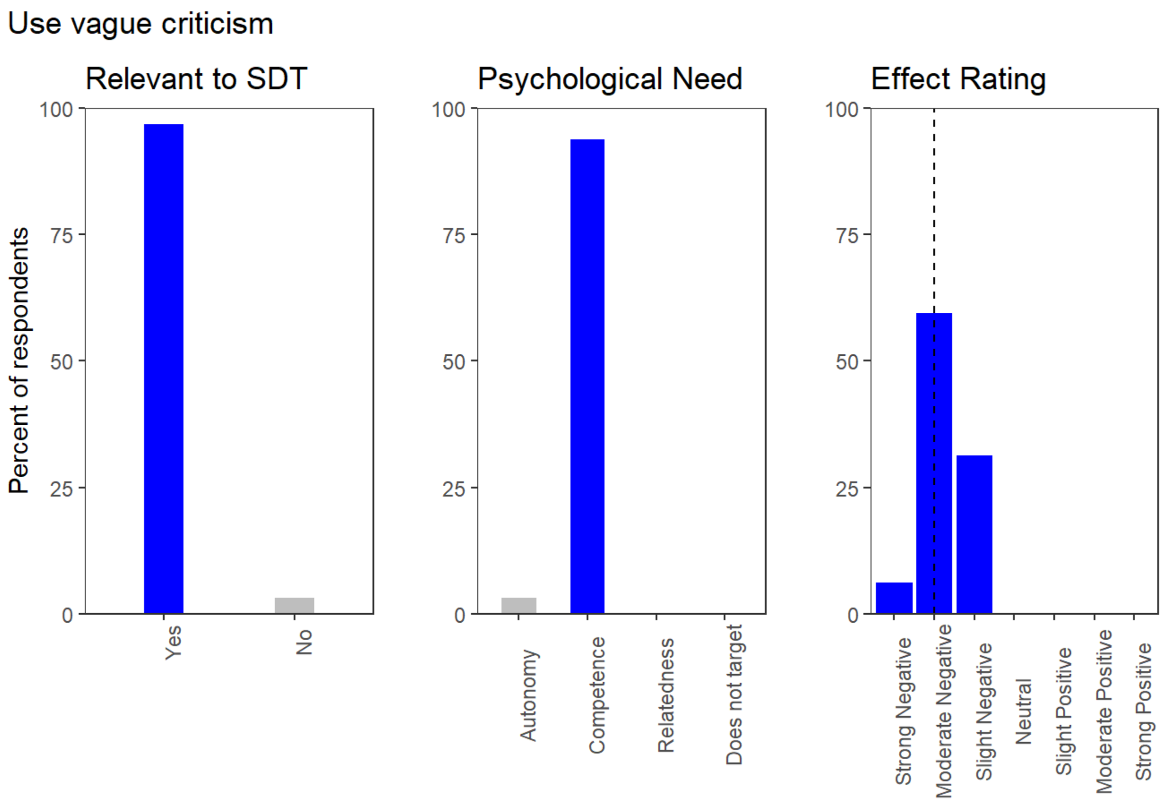
Provides vague critical feedback with no instruction of how to improve

Example Behaviour:

"Come on, James, you need to do better"

Function Description:

Creates ambiguity regarding strategies for students to increase competence



TMB#40

Exhibiting solutions or answers

Description:

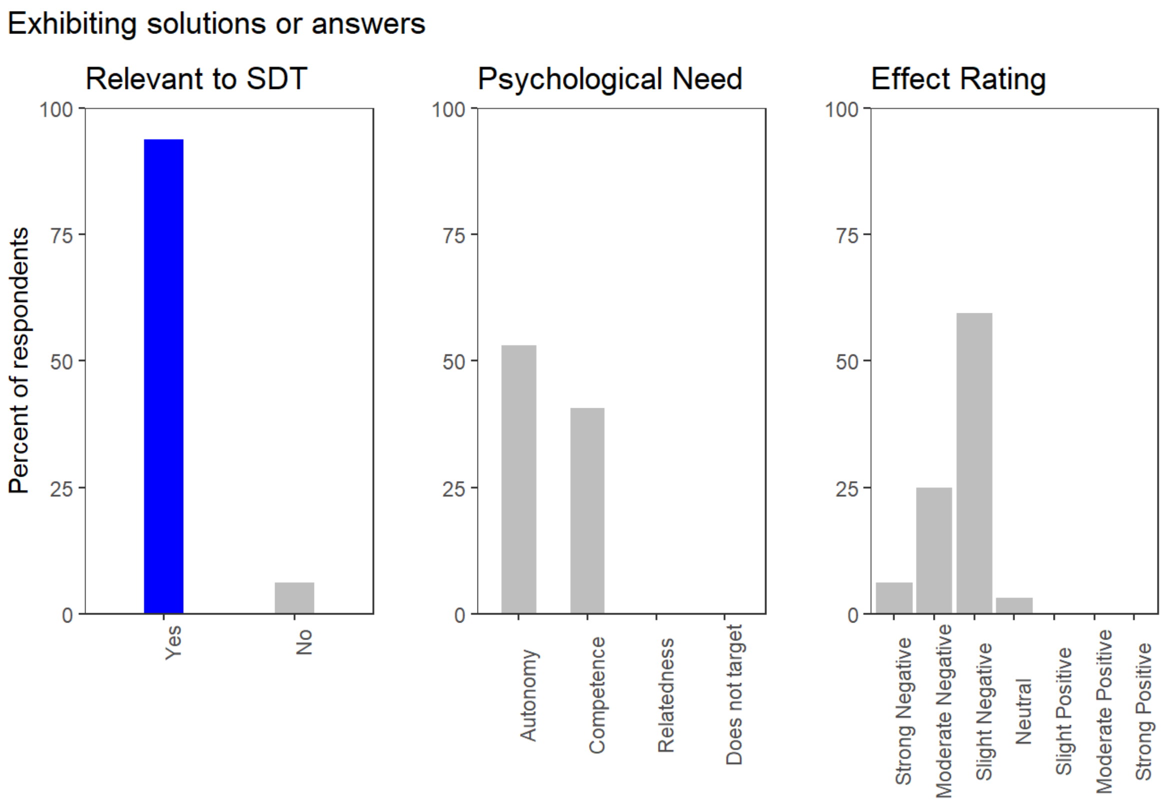
Give answers to problems instead of letting students figure it out

Example Behaviour:

"The answer is 42"

Function Description:

Stifles self-directed learning and provides external locus of causality for succes (i.e., from the teacher)



TMB#41

Praise winning via peer comparison

Description:

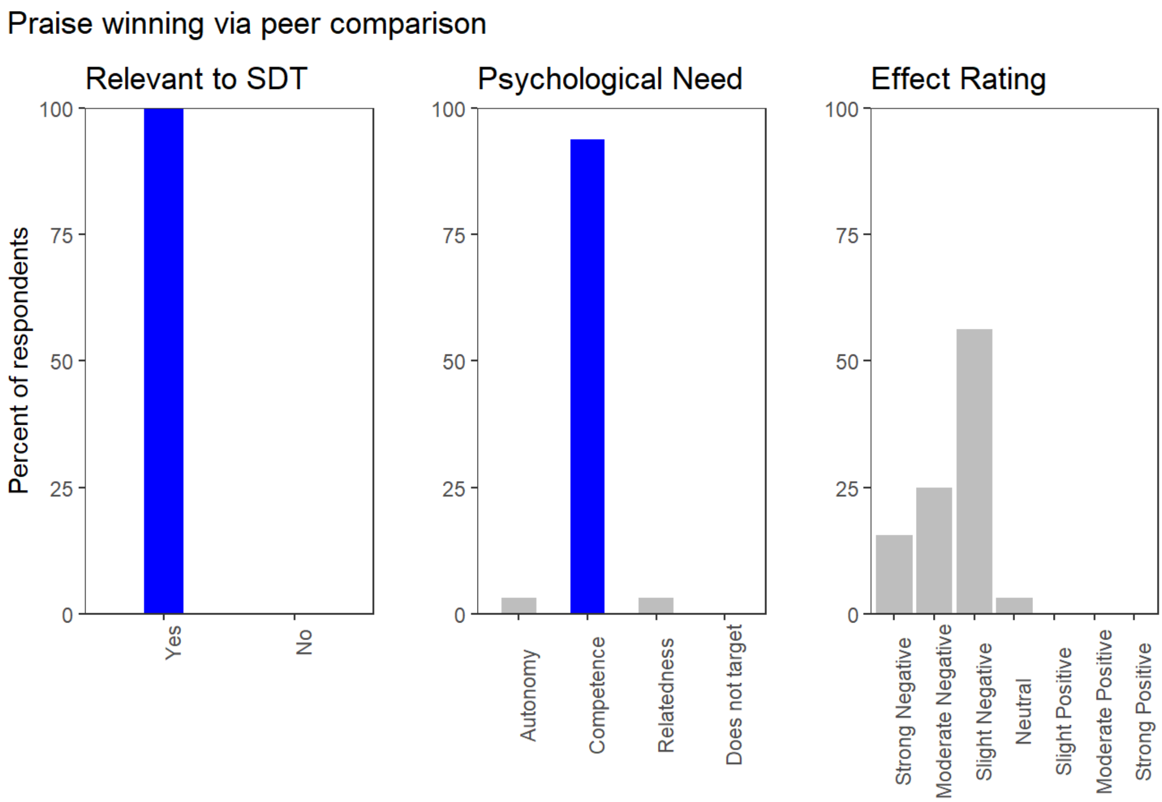
Congratulate winners so that everyone knows who did the best

Example Behaviour:

"The highest score on the exam was John"

Function Description:

Emphasises peer comparison and establishing a sense of competence, meaning few students experience success by being the best



TMB#42

Use praise as a contingent reward

Description:

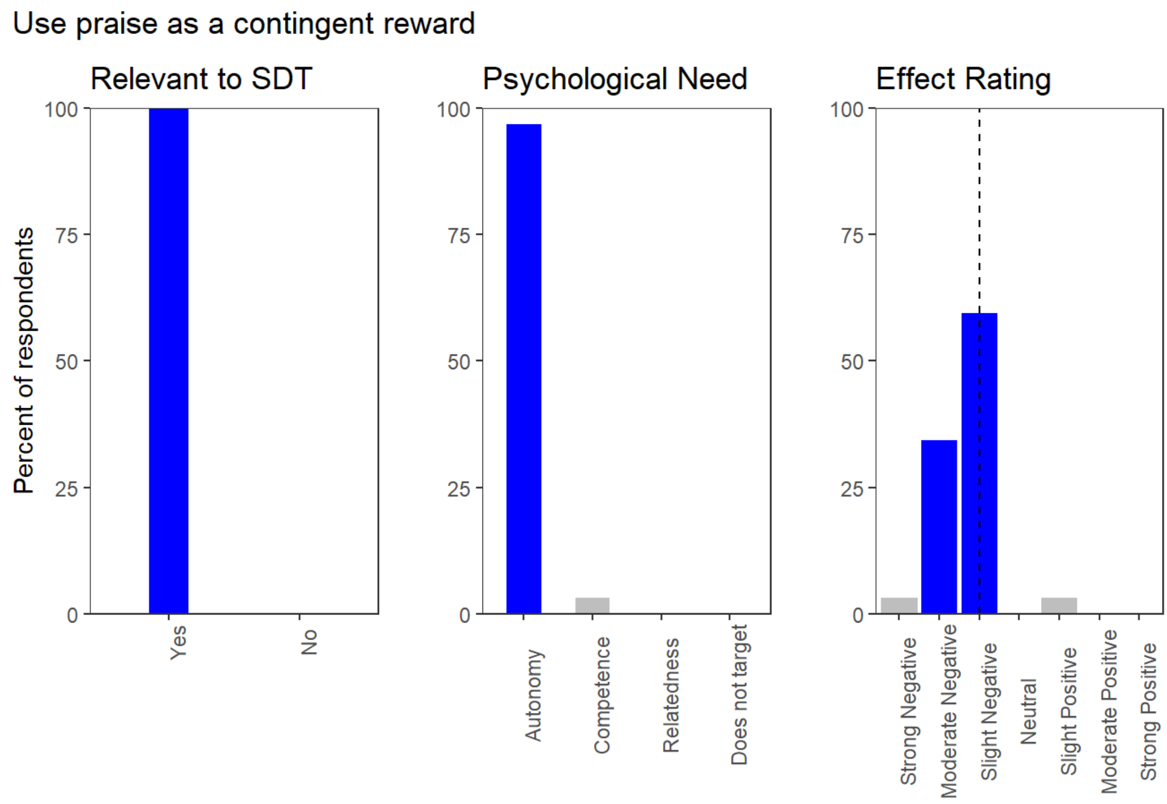
Praise students only when they do what they are told

Example Behaviour:

Teacher says to a student "Well done!" when they do what they were told

Function Description:

Increases perceived external incentives for doing an activity that is favoured by a teacher



TMB#43

Set competitive goals

Description:

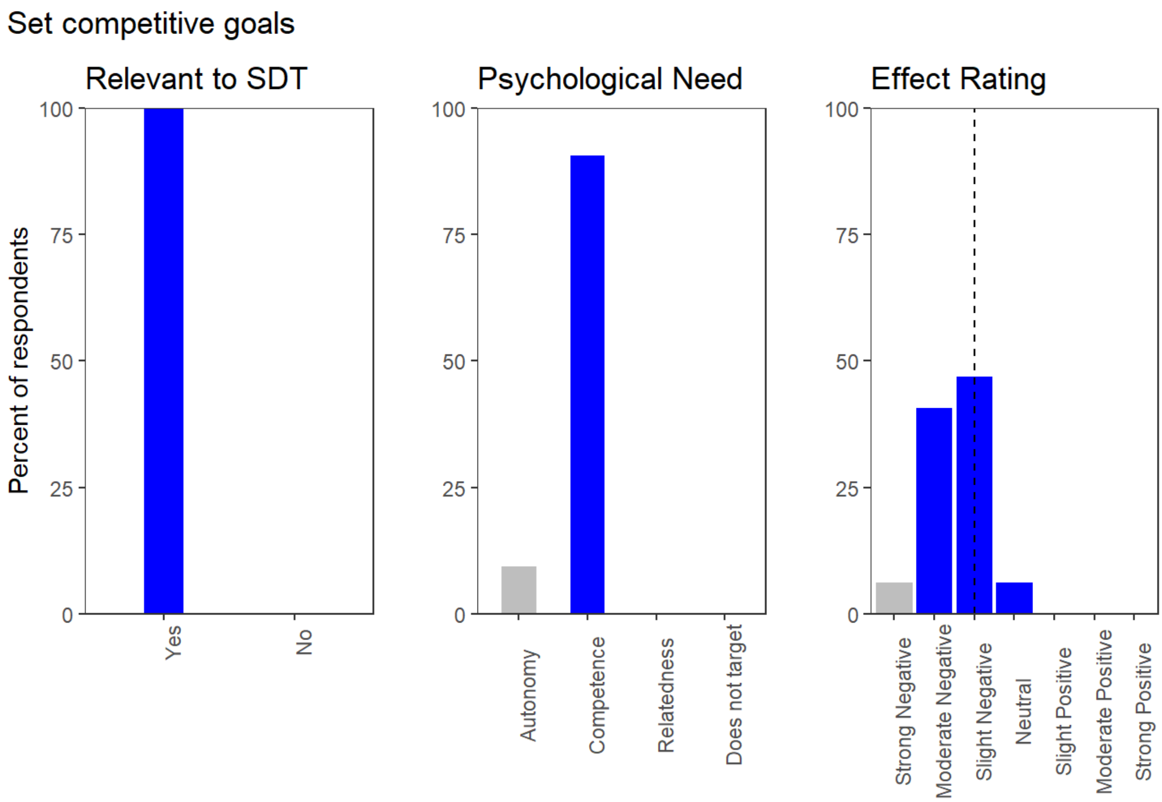
Set up activities where the goal is to do better than other students so that students compete against each other

Example Behaviour:

"whoever completes these problems in the fastest time wins"

Function Description:

Provides extrinsic reasons for working hard and few opportunities for success (i.e., winning)



TMB#44

Chaotic or Absent Teaching

Description:

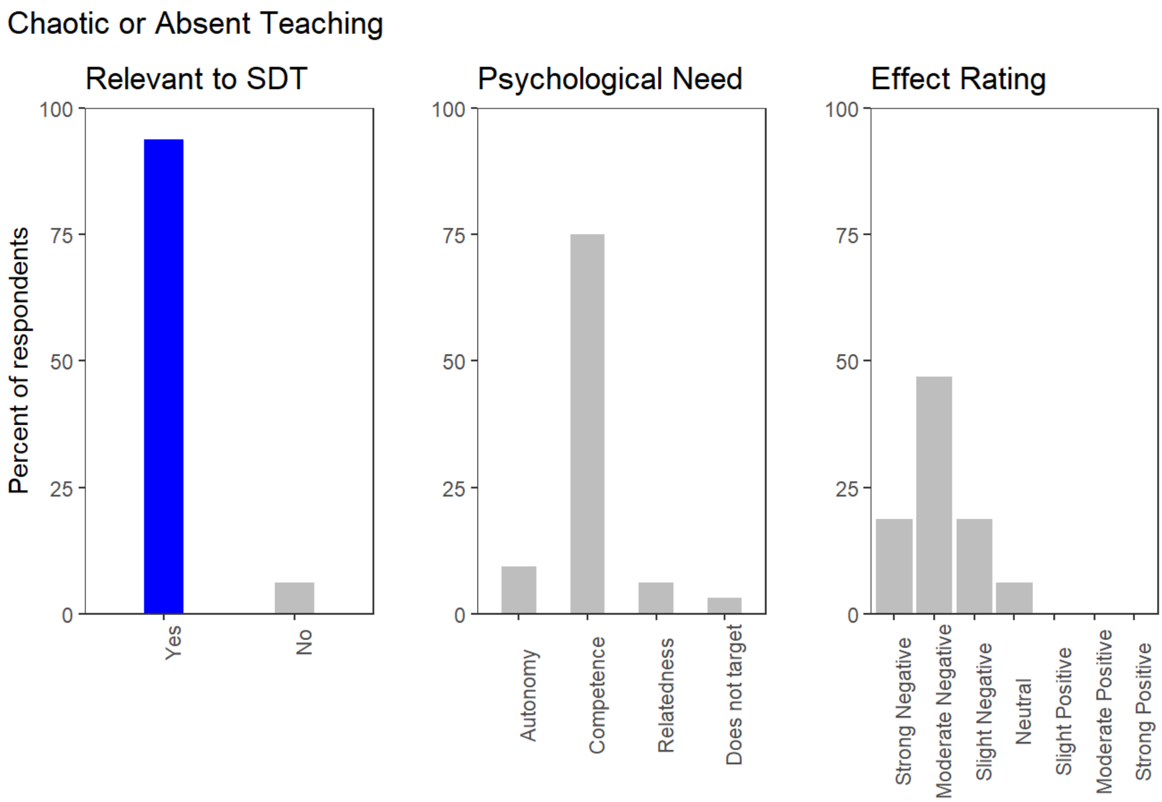
Leave students without clear instructions so the class waits or is disorganised while the teacher does something else

Example Behaviour:

Teacher leaves students waiting when arranging papers at front; Teacher gives up on providing feedback so checks his/her emails in class

Function Description:

Students do not know what they should be doing to learn and do not get any feedback or structure about how to persue goals



TMB#45

Humour

Description:

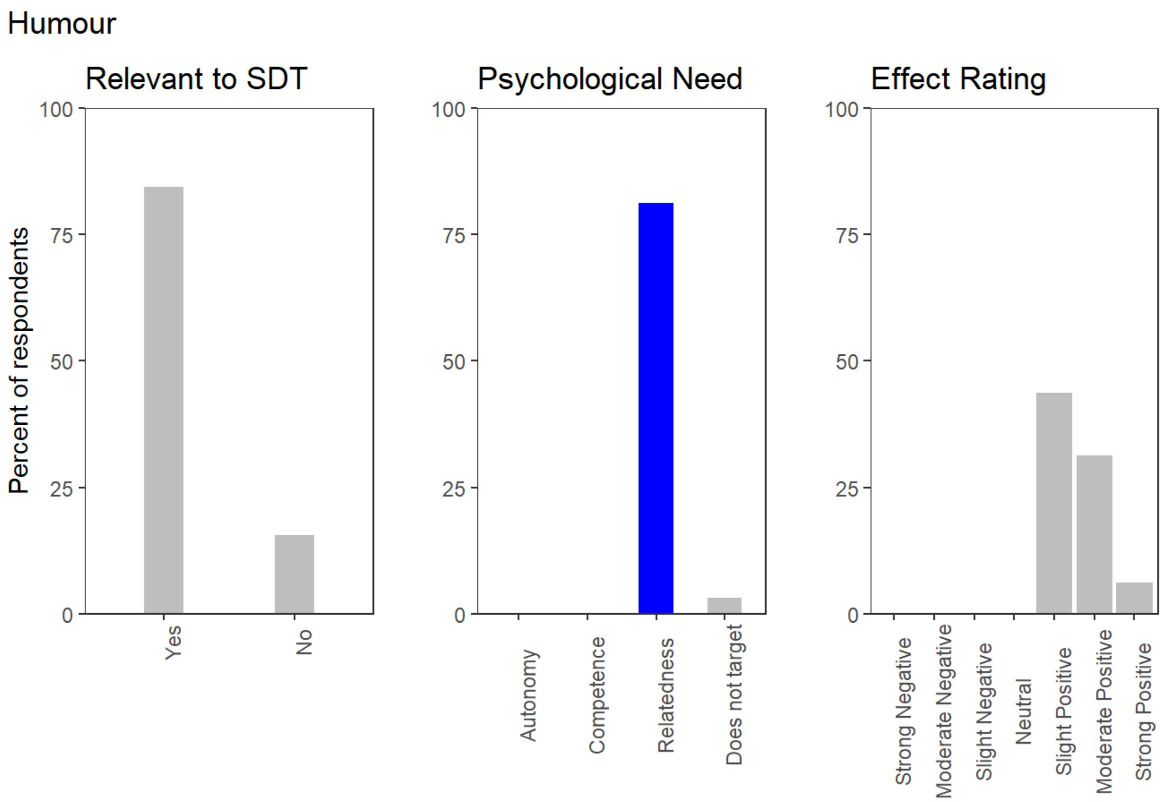
Use authentic humour so the class is fun

Example Behaviour:

"What did the triangle say to the circle? You are pointless"

Function Description:

Alleviates anxiety and reduces goal-focus; increases warmth for teacher; stimulates interest



TMB#46

Create heterogeneous groups

Description:

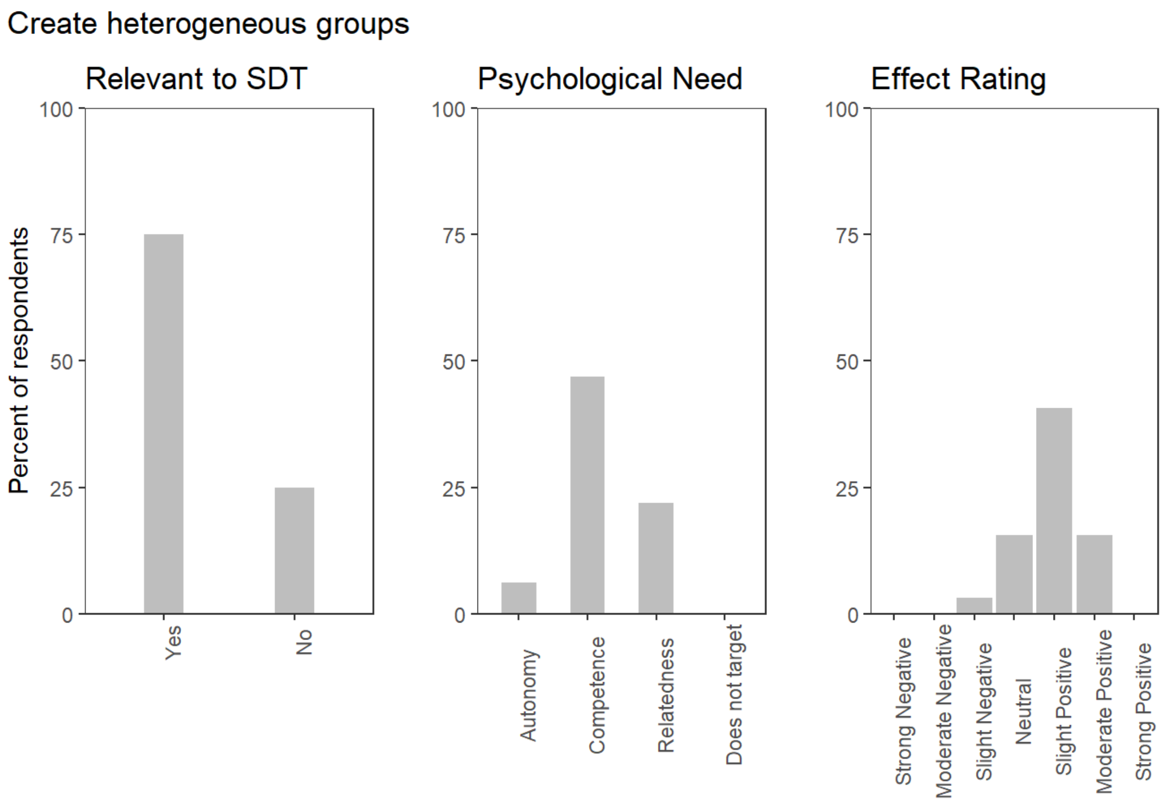
For group activities, assign students so that each group has a mix of abilities

Example Behaviour:

"Take a playing card, and find the other students with the same suit as you"

Function Description:

Removes public signalling of incompetence and ensures balanced frames of reference



TMB#47

Provide a variety of activities

Description:

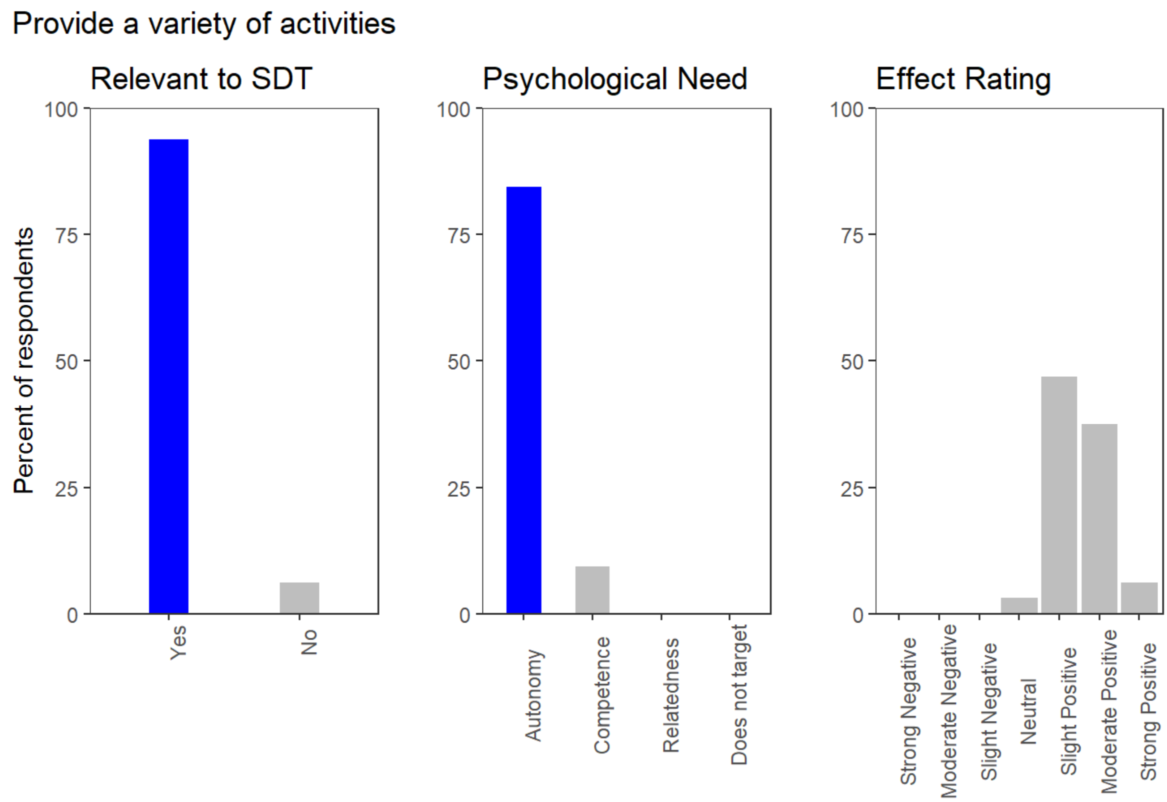
Provide a variety of activities in a way that keeps things interesting

Example Behaviour:

Teacher regularly changes the format of the class (debates one lesson, worksheets the next), and presents content in dynamic ways (teaches US History using Hamilton)

Function Description:

Reduces boredom



TMB#48

Ask questions to check knowledge

Description:

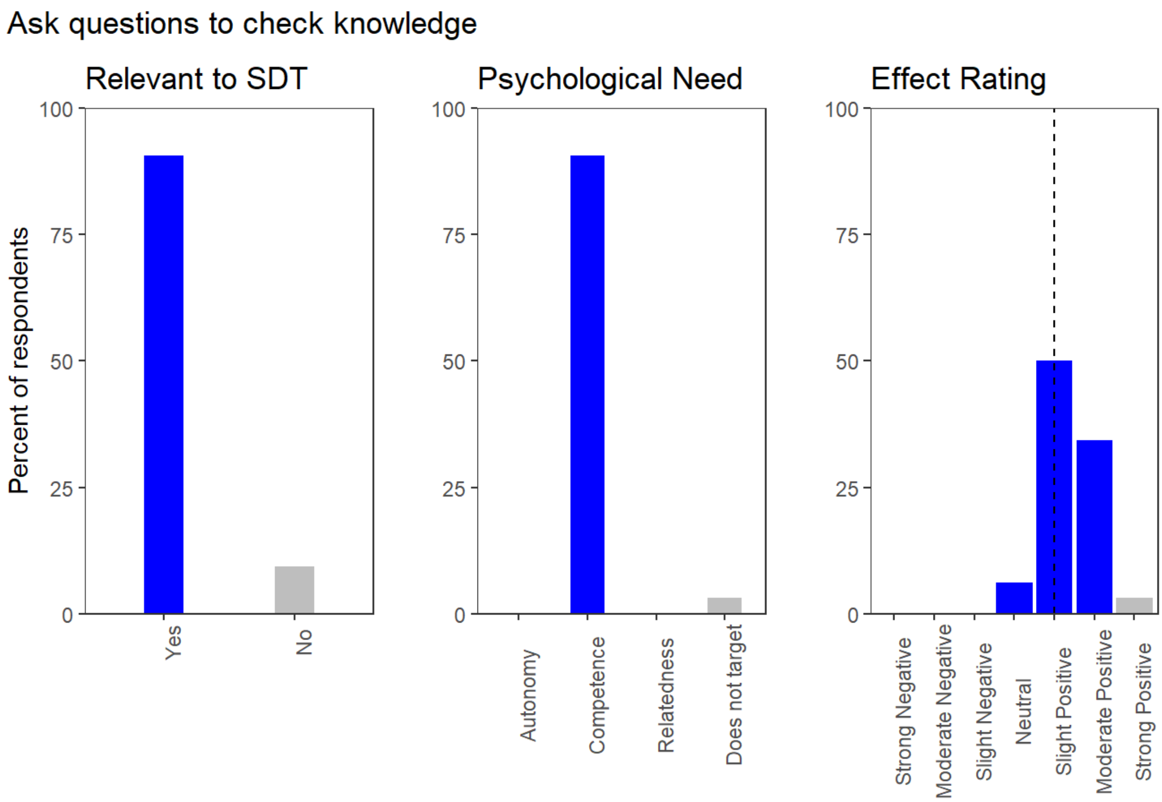
Ask the students clarification questions that check what students know

Example Behaviour:

"What is the B in BODMAS?"

Function Description:

Fosters common understanding of goal-directed behaviours



TMB#49

Ask questions to expand understanding

Description:

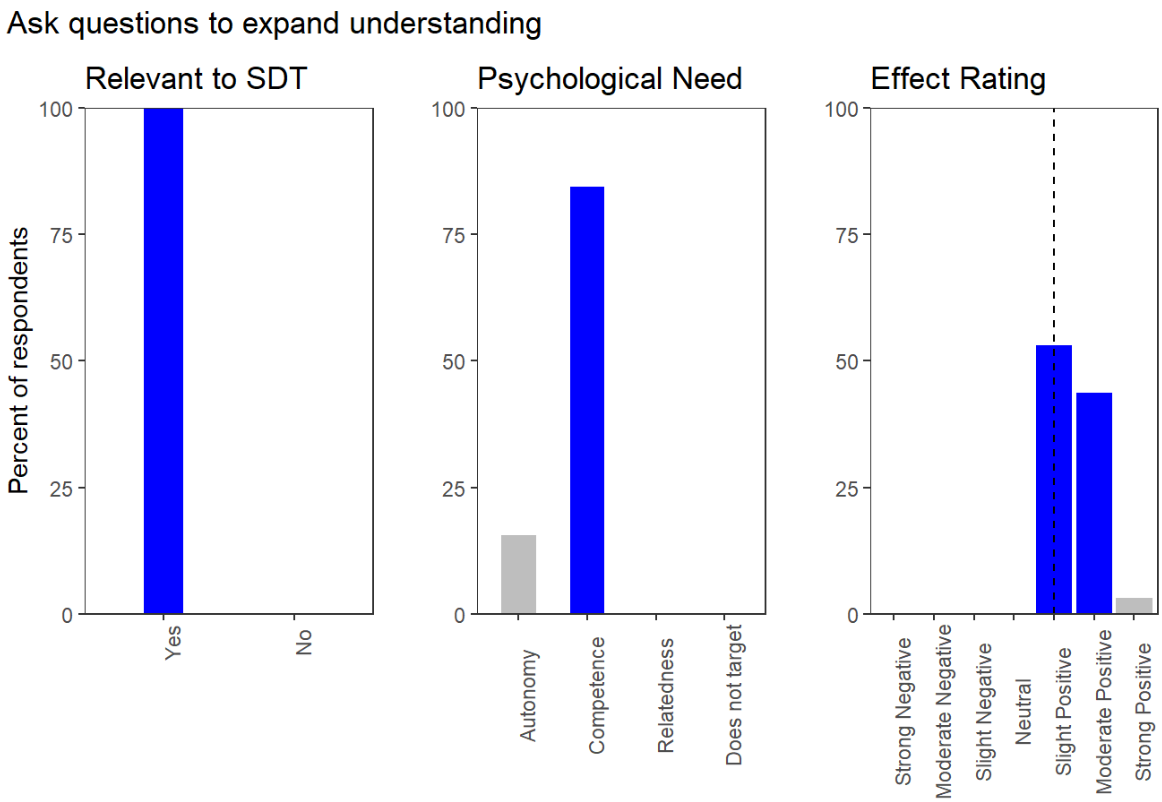
Questioning to expand understanding or thinking

Example Behaviour:

PE - "What other sports do we use these skills?", Maths - "When might we use division in our daily lives?"

Function Description:

Fosters deeper understanding of how knowledge fits together



TMB#50

Prefer open-ended questions over closed questions

Description:

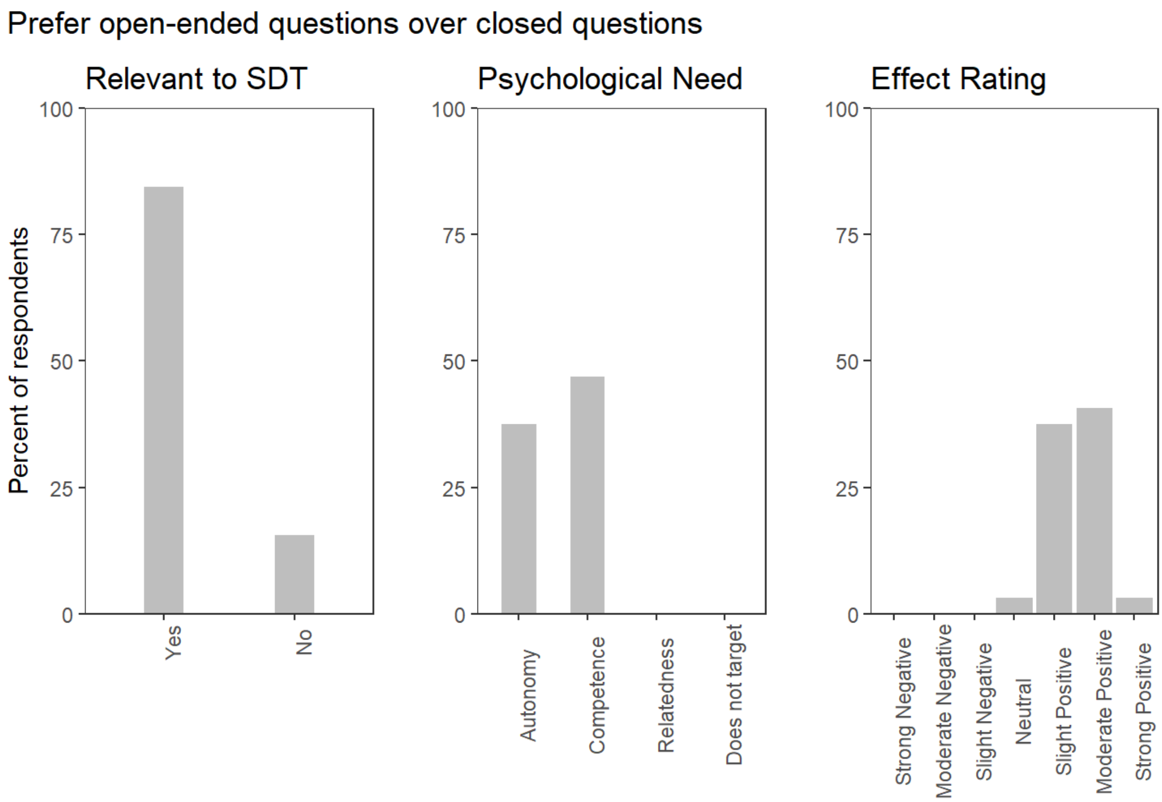
Ask questions that require many words to answer

Example Behaviour:

Ask questions starting with "why", "how", or "what" rather than "do", "is", or "are"

Function Description:

Facilitates student self-expression and deeper thinking



TMB#51

Apply fair punishments

Description:

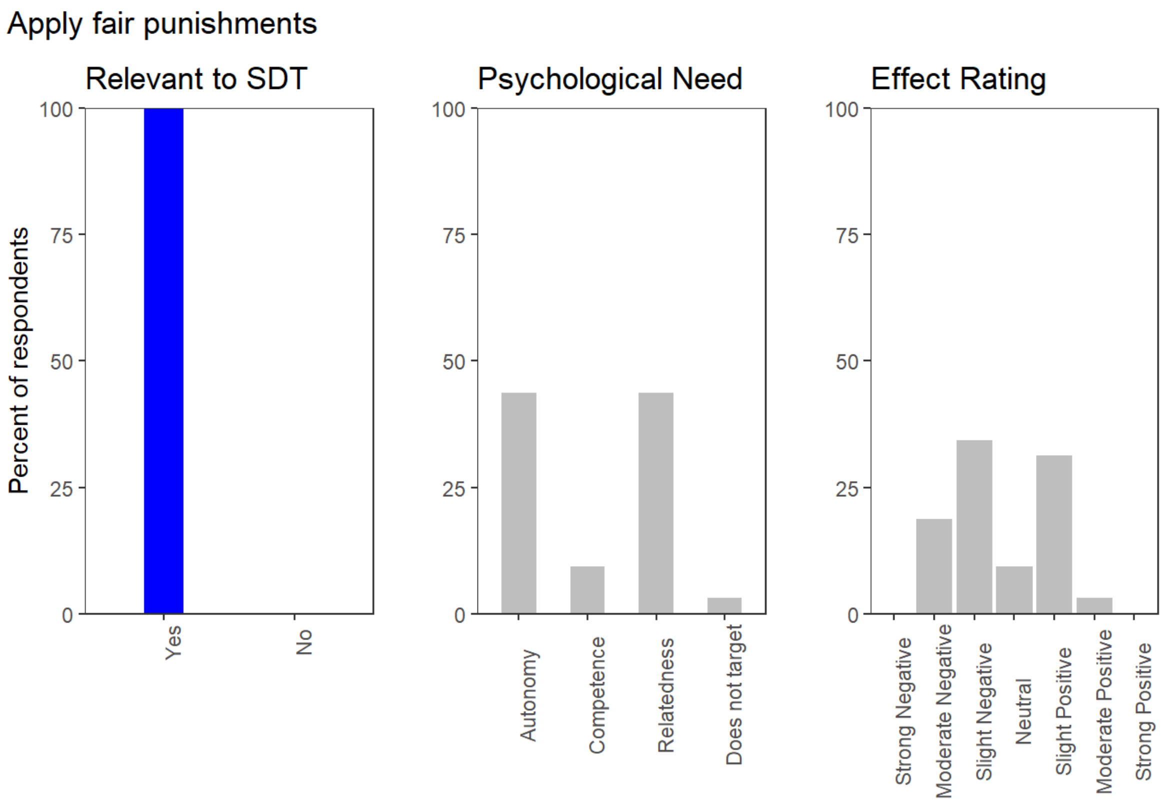
Provide punishments fairly so students who misbehave are treated equally

Example Behaviour:

Sending both of two students out of class when they misbehave or break a rule

Function Description:

Ensures misbehaviour is consistently and reliably met with external contingencies



TMB#52

Set up activities that exclude some students

Description:

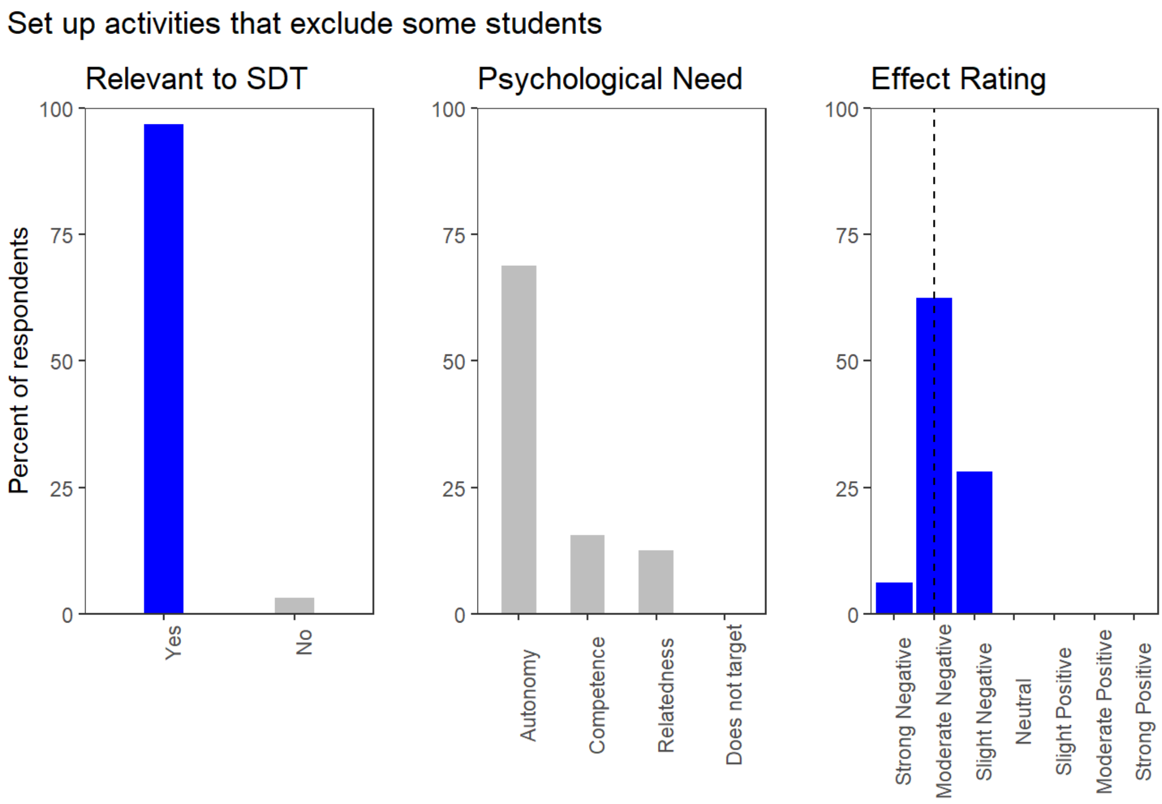
Set up activities so there are times where some students are not doing anything

Example Behaviour:

"if you have finished the questions, just sit quietly until everyone else is finished"

Function Description:

Students do not have opportunities to engage even if they want to



TMB#53

Teaching children to set intrinsic life goals for learning

Description:

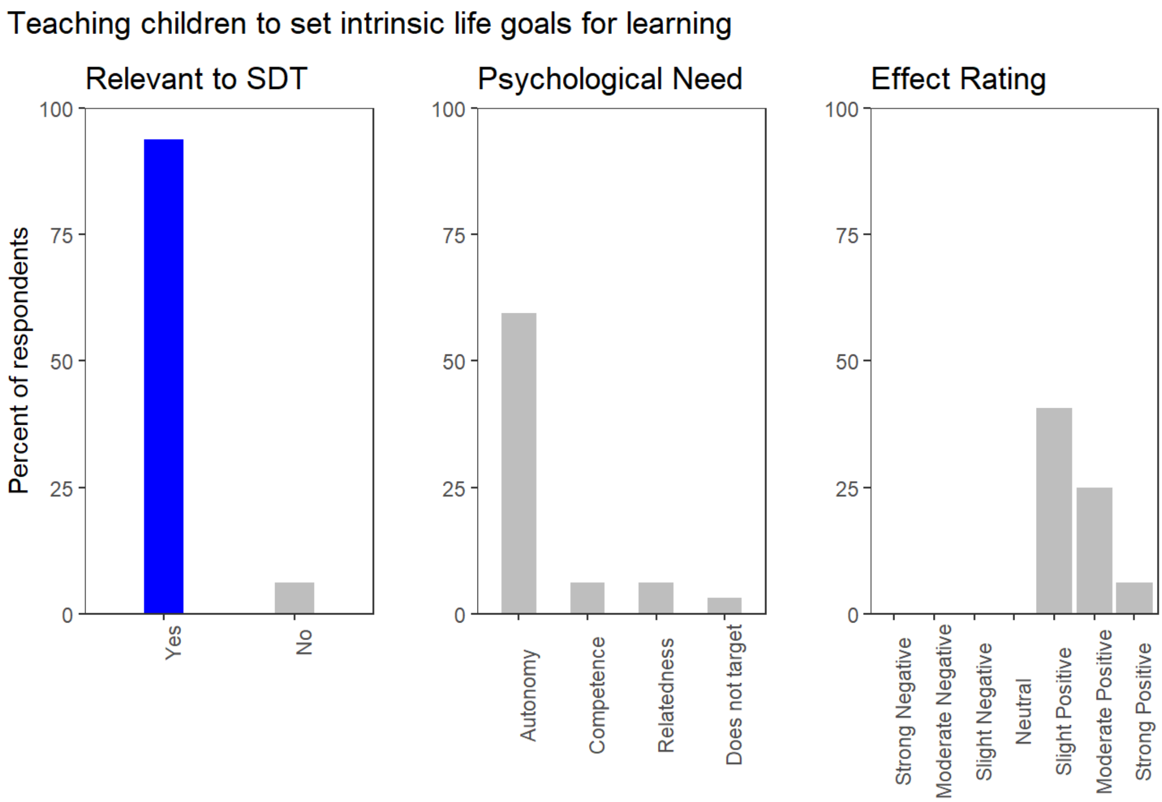
Teach students to set intrinsic life goals for learning such as "overtly healthy attitudes toward the learning process (e.g., embracing challenges, enjoyment of learning), "helping others" (e.g., how it applies to helping others or bettering one’s community).

Example Behaviour:

"Reading helps me to gain knowledge about life" or "I want to use my reading skills to read to little kids".

Function Description:

Students will try to understand the lessons more, become better at doing the activities, so that students can help others someday, or discover something interesting.



TMB#54

Use parables, stories, analogies, or metaphors

Description:

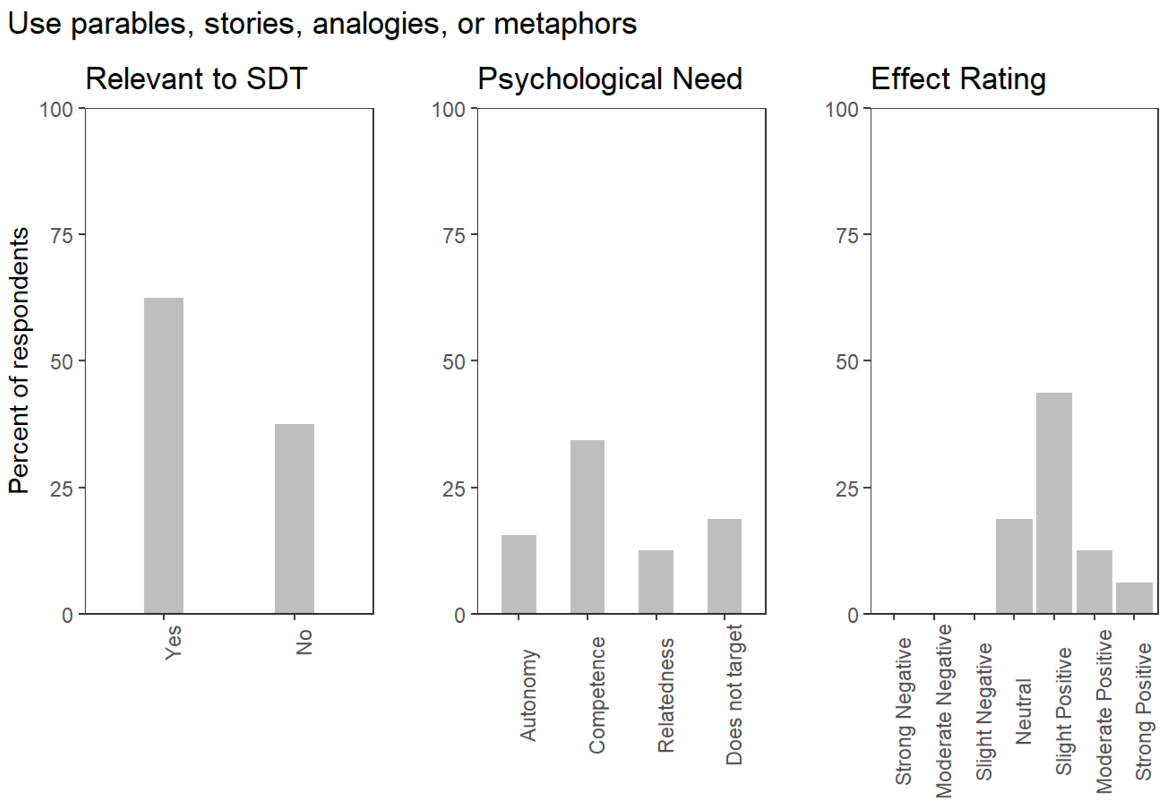
Use parables, stories, analogies, or metaphors to help students to connect abstract constructs into concrete examples

Example Behaviour:

"mistakes are stepping stones, not stumbling blocks"; "treating others well is like sowing good seeds, eventually you'll reap a good harvest"

Function Description:

Promote empathy through narratives that students can connect to, and provide examples that students could follow.



TMB#55

Adopting student initiatives

Description:

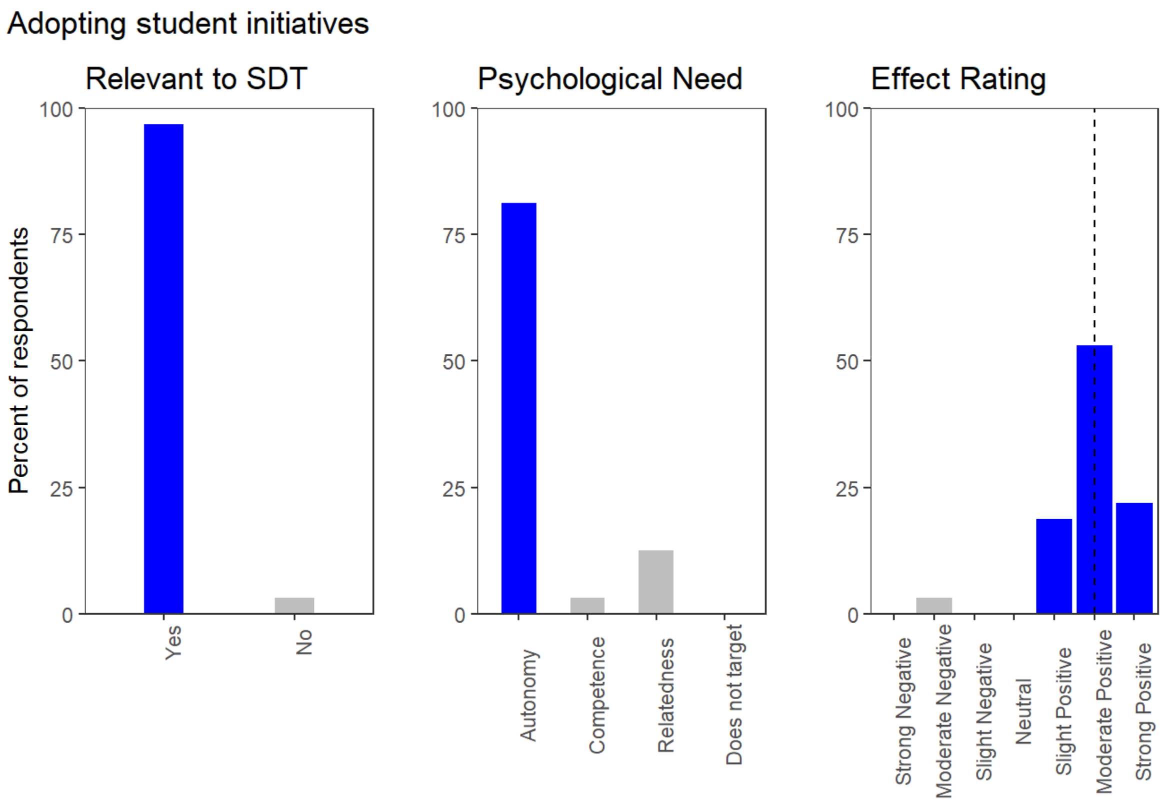
Take student suggestions into learning activities when they arise

Example Behaviour:

"That's a great idea. We can do that activity in this session."

Function Description:

Encourages and rewards student initiative and self-management of learning.



TMB#56

Set goals on behalf of the class

Description:

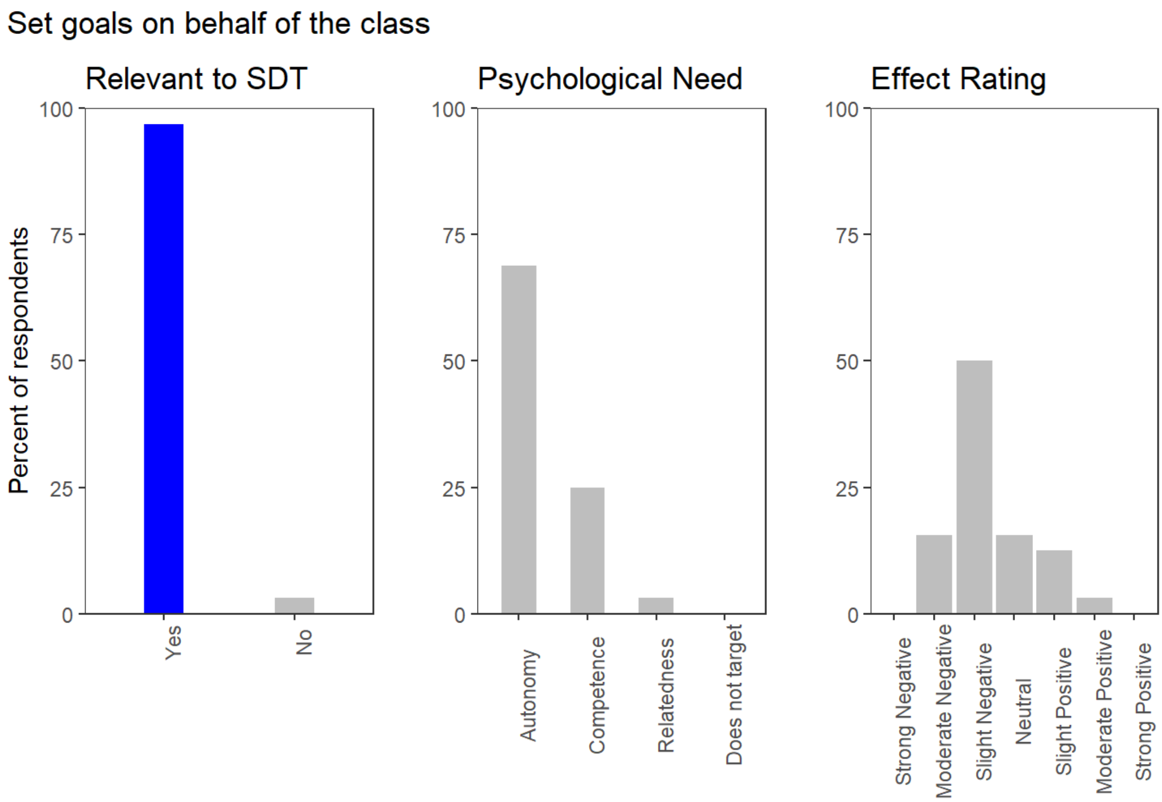
Teachers setting expectations for students rather than letting them to decide their own goals

Example Behaviour:

"The goal for tonight is to complete the all activities on page 4."

Function Description:

Provides extrinsic reasons for the goals and for doing activities and the student autonomy is undermined



TMB#57

Make mistakes or give incorrect feedback

Description:

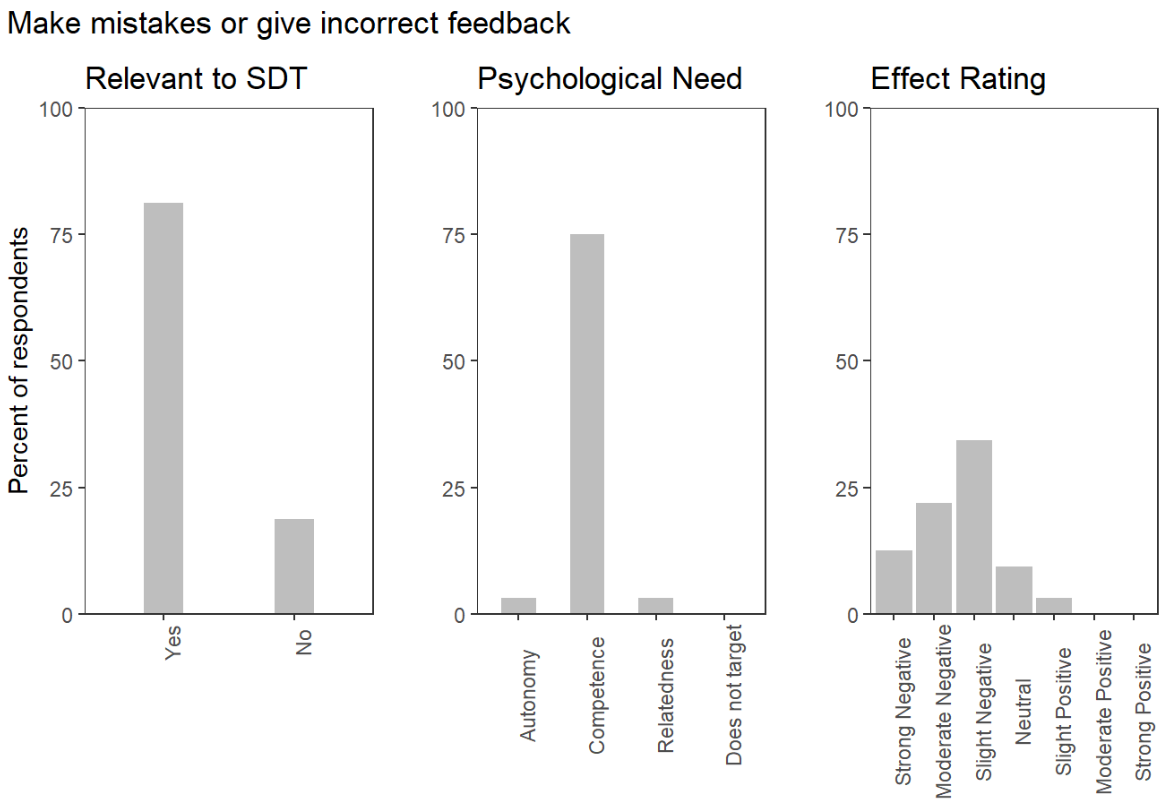
The teacher models in a way that is unlikely to produce the desired outcome, without providing the correct approach.

Example Behaviour:

"Notice how I try to catch with my arms straight in the air, pointing toward the ball."

Function Description:

Undermines progress because students are likely to reproduce the undesirable approach.



TMB#58

Group students with similar interests

Description:

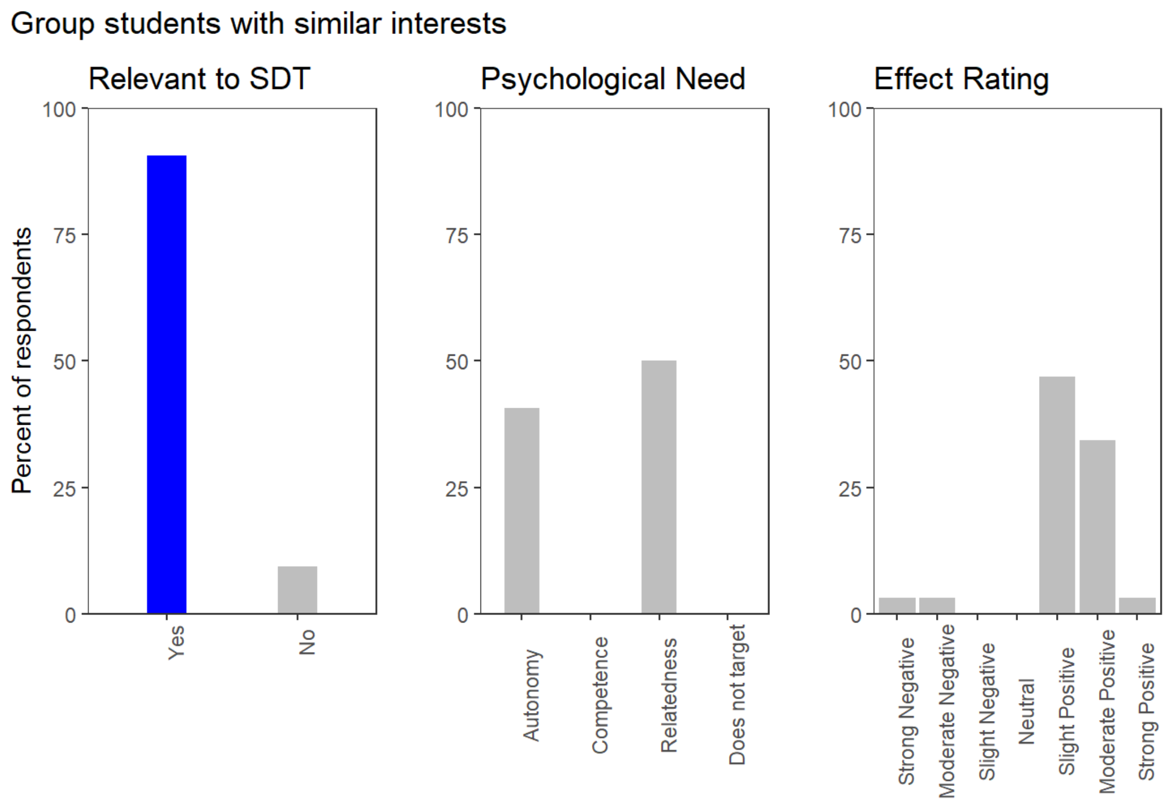
Create groups in the class where students with similar values or interests can work together on problems

Example Behaviour:

When studying geography, grouping musical students to look at a country's music, the sporty students to look at the country's sports, and other students to look at the country's key historical events.

Function Description:

Allows students to work on tasks—and with people—that match their interests and values.



TMB#59

Regular communication with parents

Description:

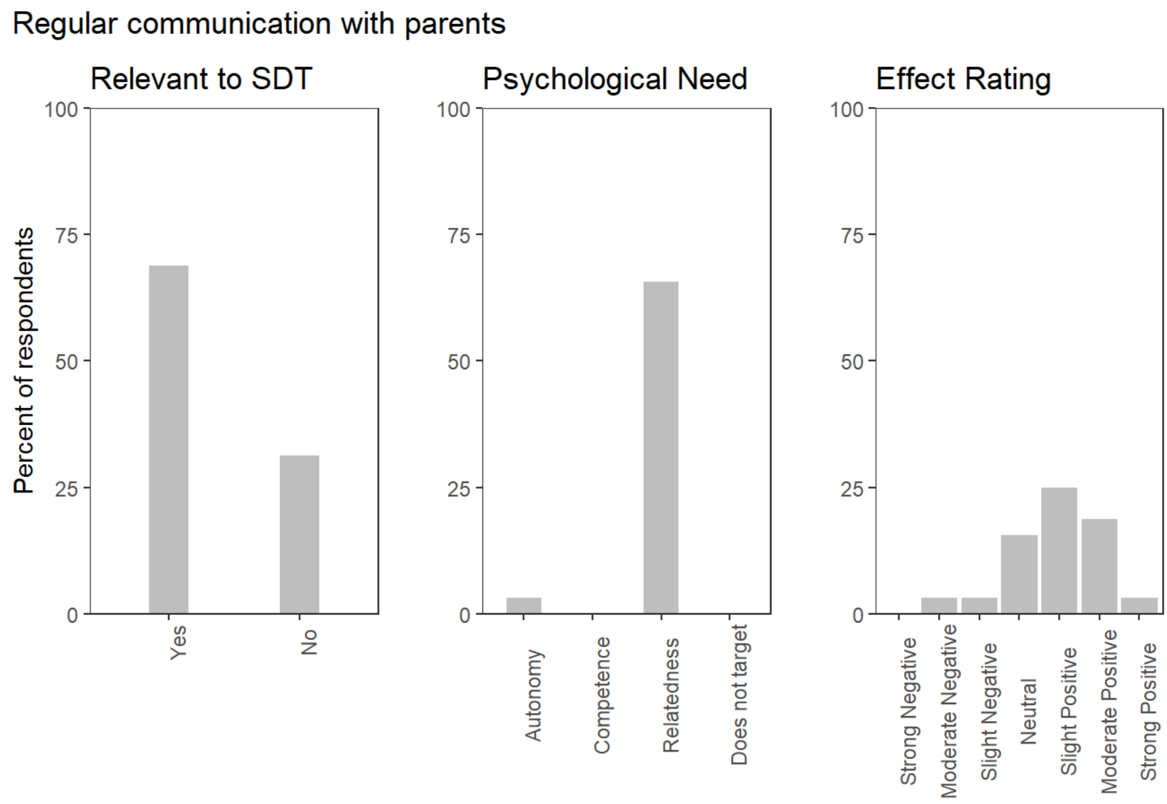
Teachers engaging in regular contact (e.g., phone, email, text) with parents about the activities in the class or of their children

Example Behaviour:

The teacher calls a parent when she notices that a student has been particularly disengaged and unenthusiastic to talk

Function Description:

Supports connections between the students' home and school life, identifying ways that key people in each domain (e.g., parents, teachers) can support each-other



TMB#60

Be sarcastic

Description:

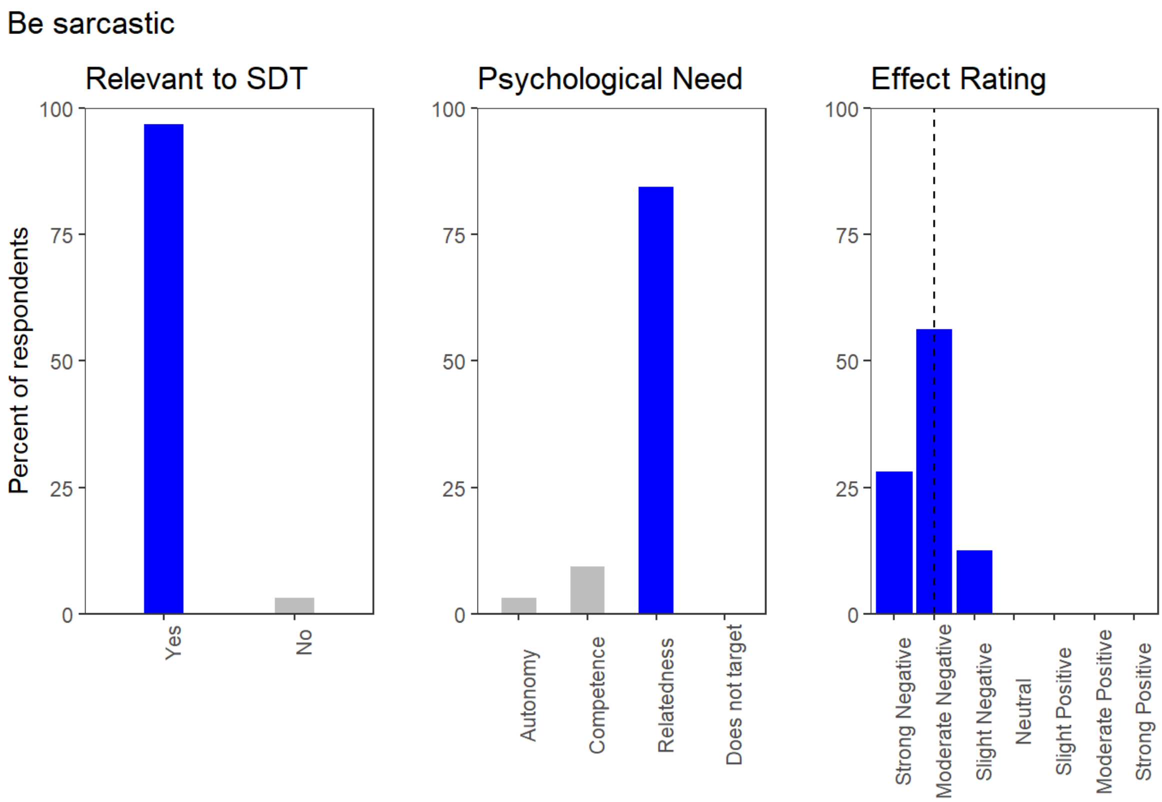
Use sarcastic negative phrases

Example Behaviour:

“Class started 3 minutes ago. Soooo nice of you to join us.” Or, “It's not like what we are learning today is important or anything.”

Function Description:

Reduces student self-esteem and devaluates their sense of being a valuable person and students



TMB#61

Unfair use of praise

Description:

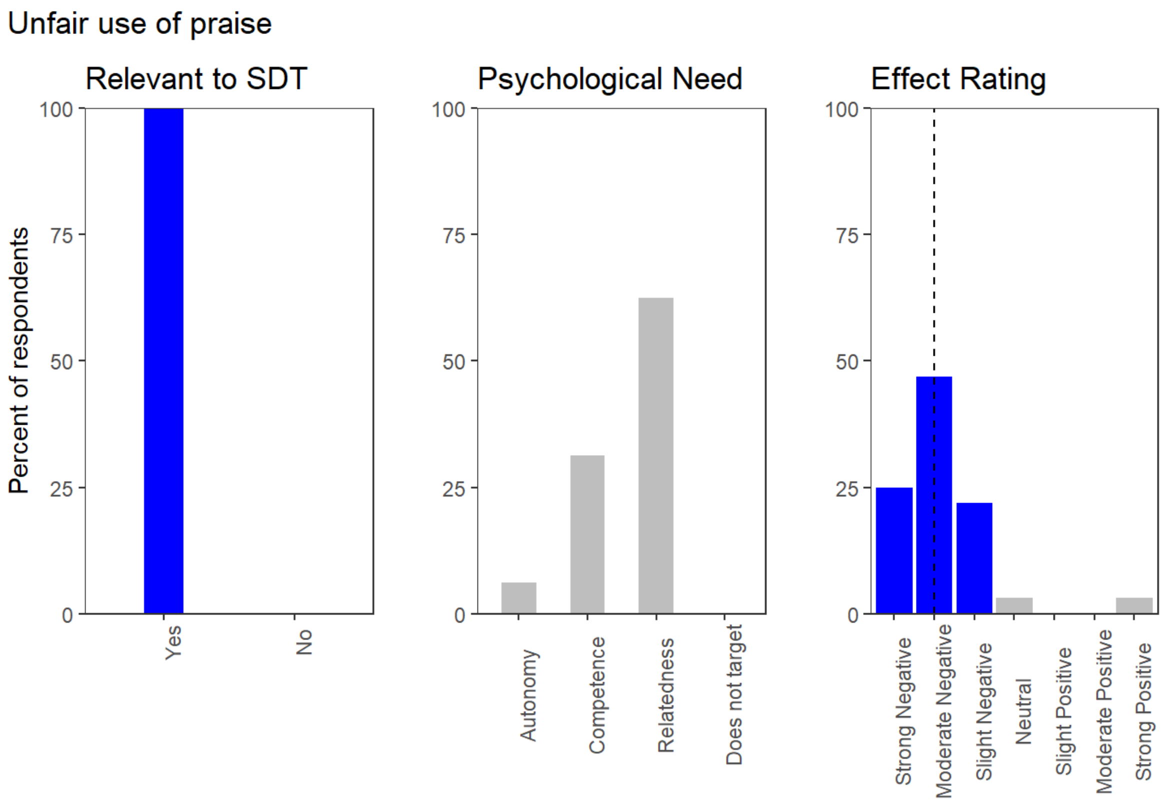
Praises students unfairly or unequally; shows favourites

Example Behaviour:

Complementing only one of three students who completed a problem a creative way

Function Description:

Makes students feel like some are more worthy of praise than others



New TMBs suggested by the experts

TMB#62

Ask students about their experience of lessons

Description:

Ask students for feedback about how classes are going; could apply to either the content of lessons or the process/learning design

Example Behaviour:

"On these sheets, please write down what you liked about today's lesson, what you didn't like, and what was most unclear. Remember it's anonymous."

Function Description:

Gives students a safe opportunity to suggest constructive input and shape the way classes are run, so lessons can better cater to their needs and interest

Appendix C.3

Delphi Round 3 Results with Plots

TMB#1

Active Learning

Description:

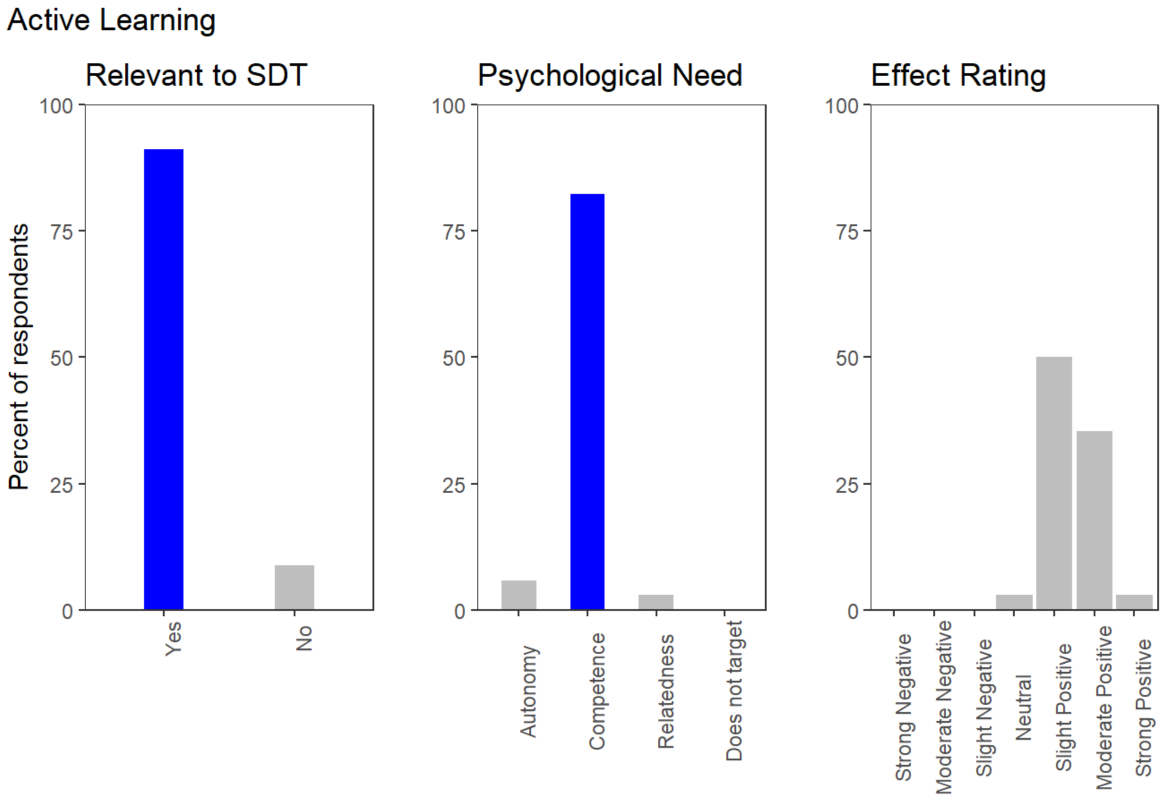
Set up activities where all students are engaged in a learning activity

Example Behaviour:

"Complete this worksheet individually to figure out how heavy the Sydney Harbour Bridge is"; "Try to make a sentence using as few of these phonemes as possible"

Function Description:

Allows each student hands-on practice with an activity designed to progress development of a skill



TMB#2

Discuss class values

Description:

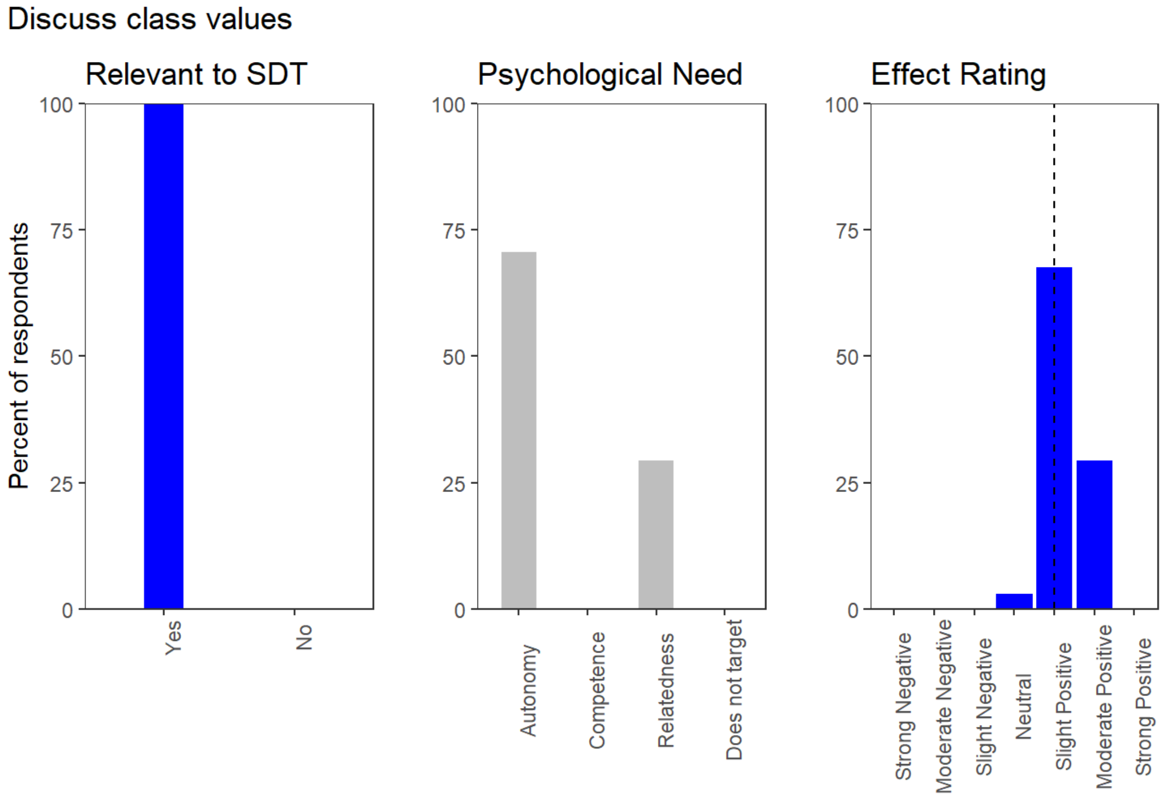
Collaboratively establish the values important to display in the class, or remind students of the collaboratively derived values

Example Behaviour:

"We all thought working hard was important, so even though many find this task difficult, see if you can push through to the end."

Function Description:

Connects the activities that take place in class with values that the student cares about



TMB#3

Modelling resilience by expressing vulnerability

Description:

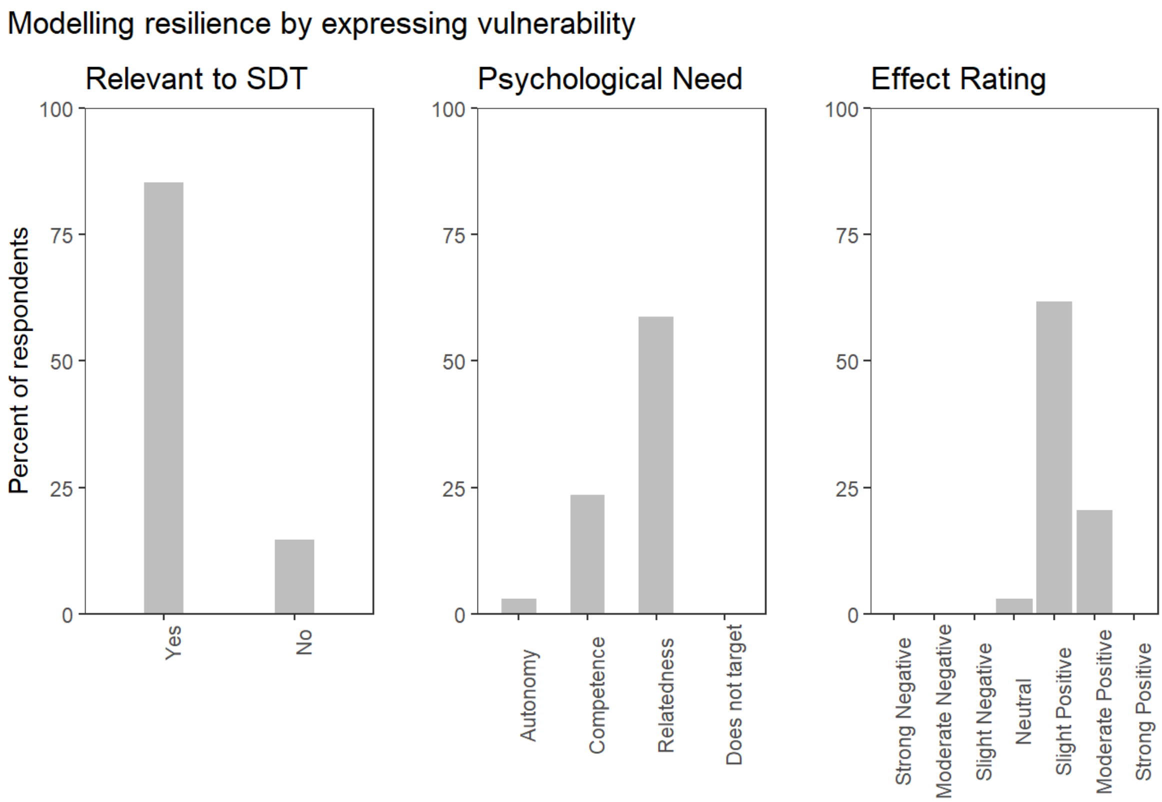
Showing that it is possible to adapt and achieve despite difficulties now or in the past

Example Behaviour:

"I struggled to write clearly for years, but I kept asking for feedback and got better."

Function Description:

Helps students to perceive the teacher as a model for coping with challenges; makes classroom more accepting space for failure



TMB#4

Teacher enthusiasm

Description:

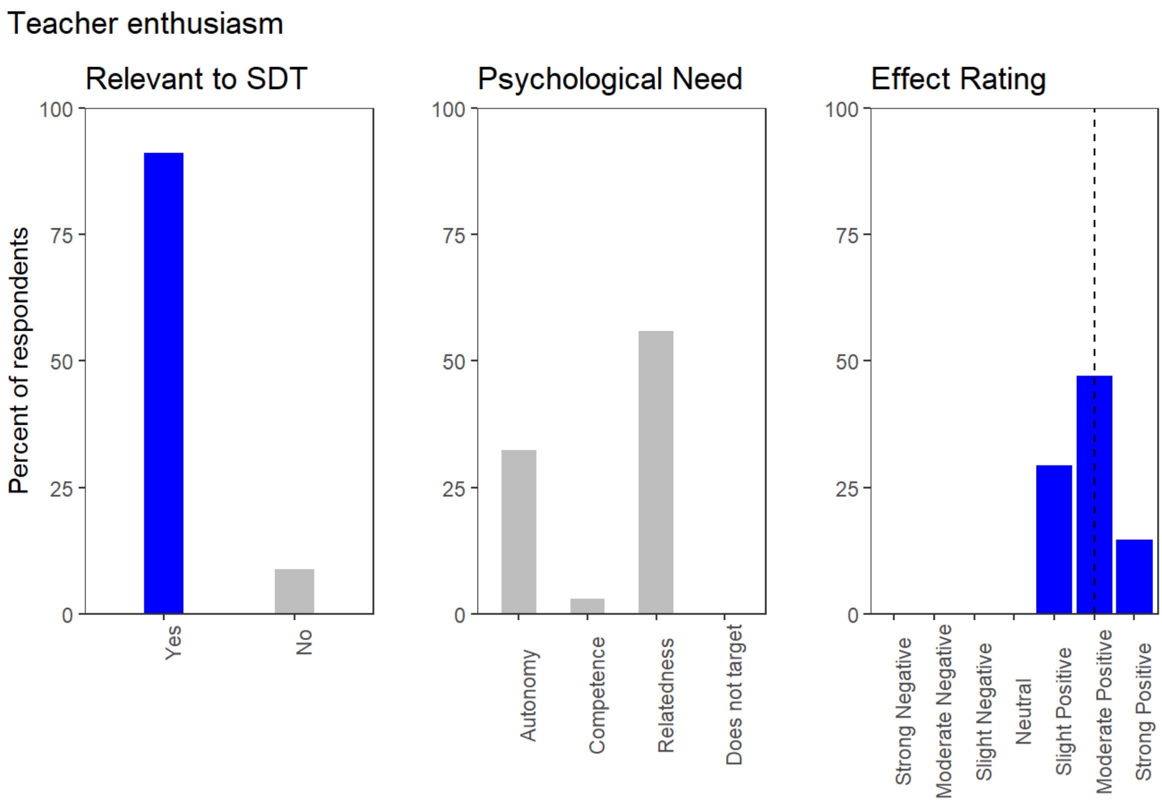
Present content enthusiastically to make things fun and interesting

Example Behaviour:

"Now I think this next part of the lesson is really interesting!"

Function Description:

Models the attitude and energy that the teacher would like the students to demonstrate; shows interest in the material.



TMB#5

Offer rewards

Description:

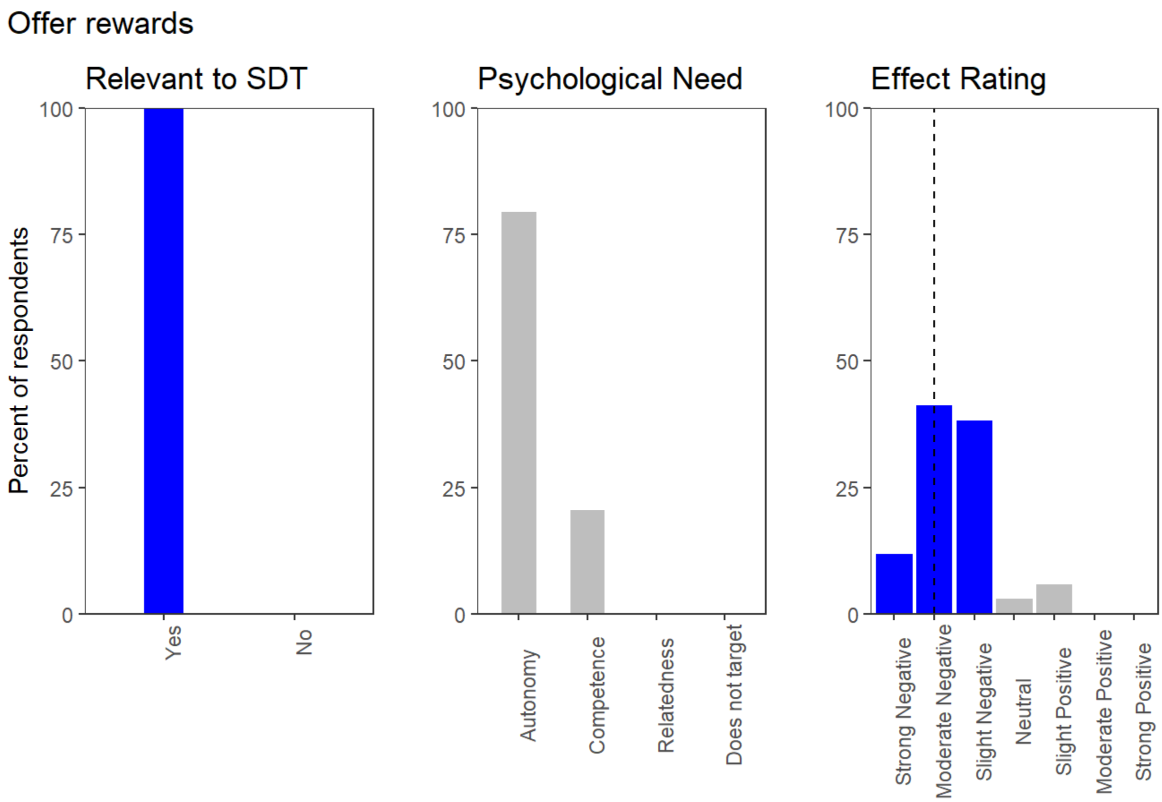
Offering—but not yet providing—extrinsic rewards: privileges or items that are not inherent to the task, but are provided in an effort to promote a behaviour.

Example Behaviour:

"If you all finish the questions, I'll play a short video clip."

Function Description:

To direct behaviour so students know what behaviour the teacher wants to see



TMB#6

Provide frequent constructive criticism

Description:

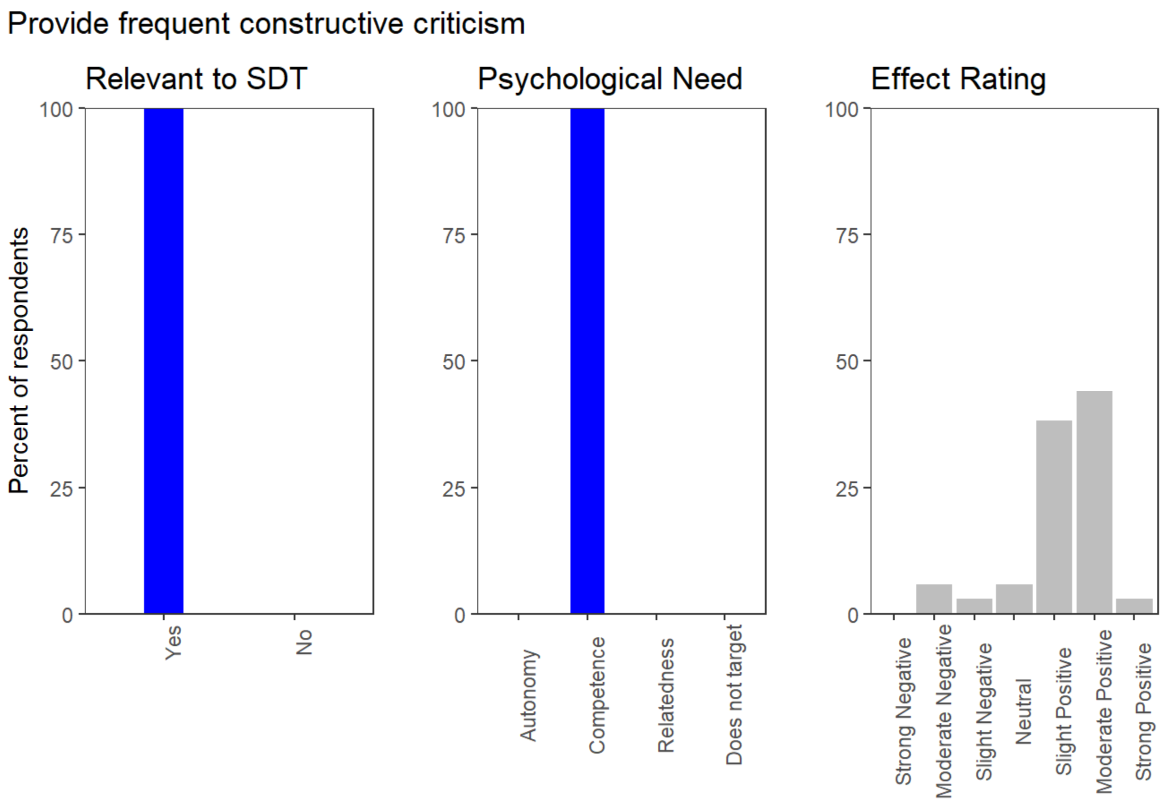
Frequently provide constructive criticism (informative feedback regarding areas of improvement)

Example Behaviour:

Consistently observing the class to provide feedback for getting unstruck

Function Description:

Promotes continual improvement in abilities.



TMB#7

Offering hints

Description:

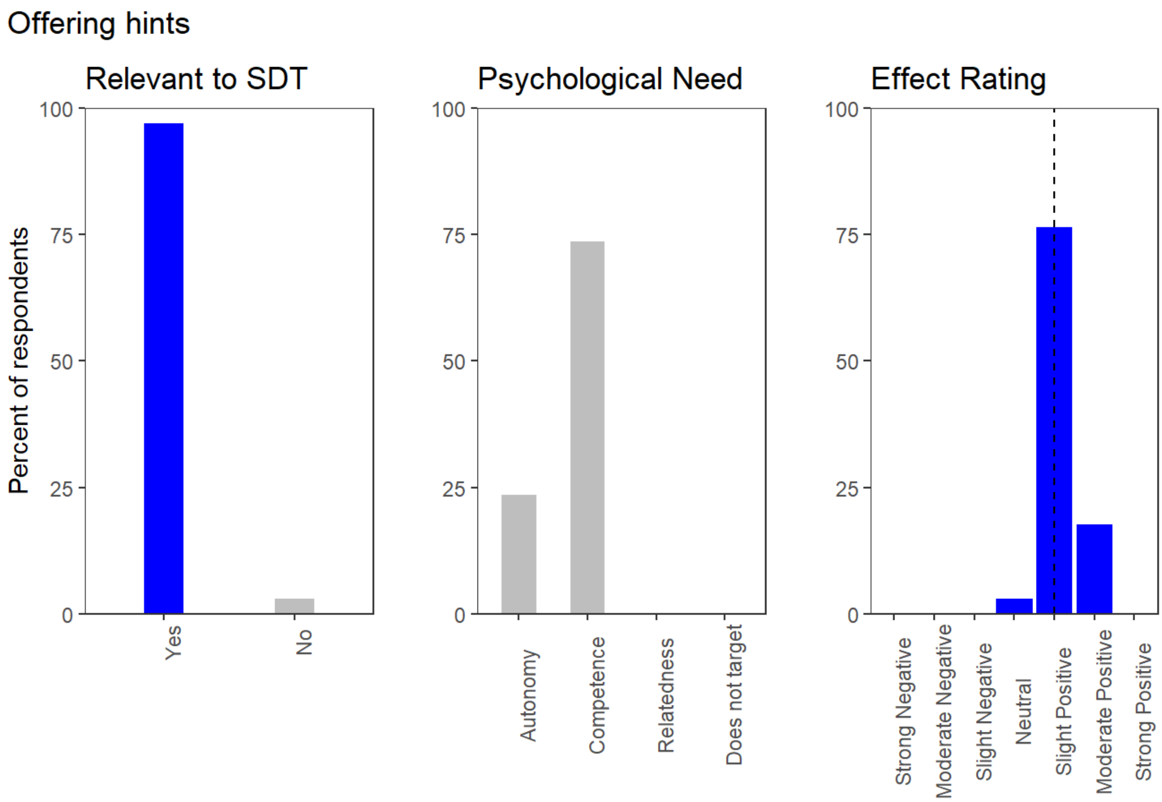
Give hints to help students along without giving them the "right answer"

Example Behaviour:

“It might be easier to start with this formula.”

Function Description:

Supports the student’s own learning processes. Allows students to maintain an internal locus of causality during learning.



TMB#8

Praise a student's fixed qualities

Description:

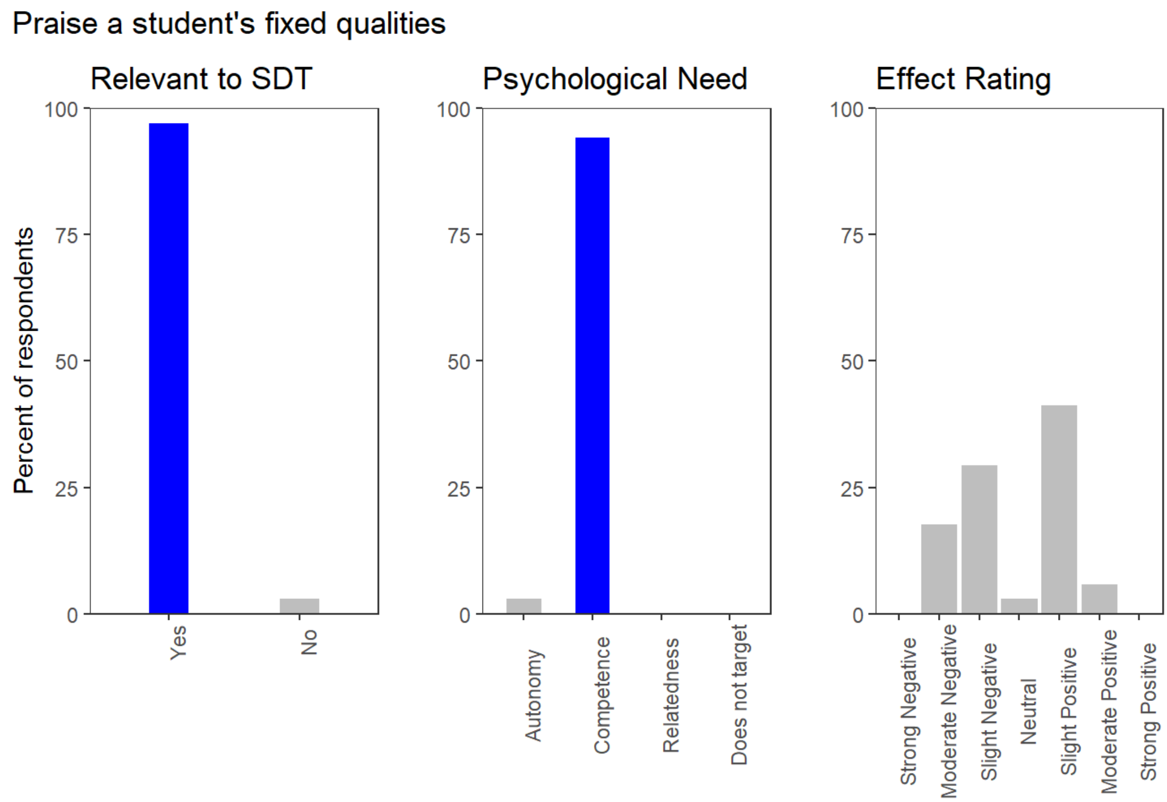
Provides praise that targets the talents or fixed qualities of students

Example Behaviour:

"You are very smart"

Function Description:

Affirms students natural abilities



TMB#9

Praise a student in public

Description:

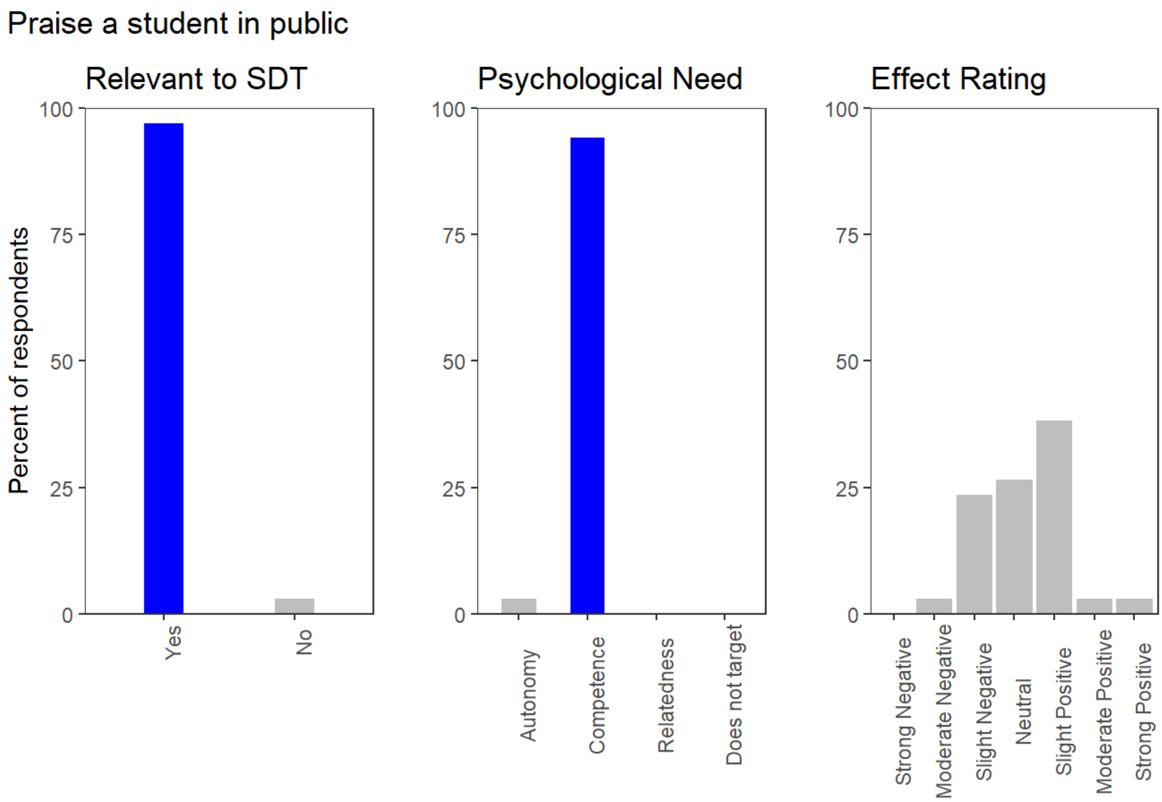
Praise a student in public

Example Behaviour:

Praise in front of the class

Function Description:

Generates pride within students receiving praise



TMB#10

Provide frequent praise

Description:

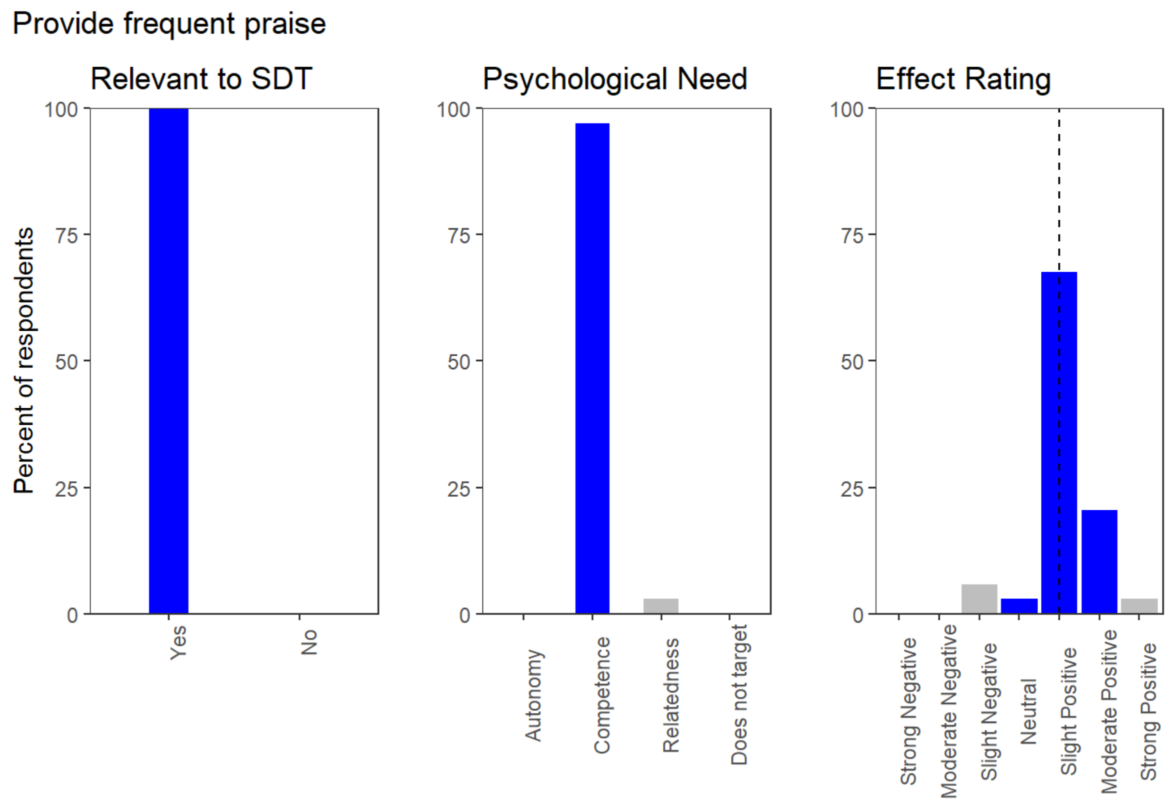
Frequently praise students

Example Behaviour:

Teacher consistently observes the class, praising students for correct answers

Function Description:

Provides continual affirmation of progress and improvement



TMB#11

Provide extra resources for independent learning

Description:

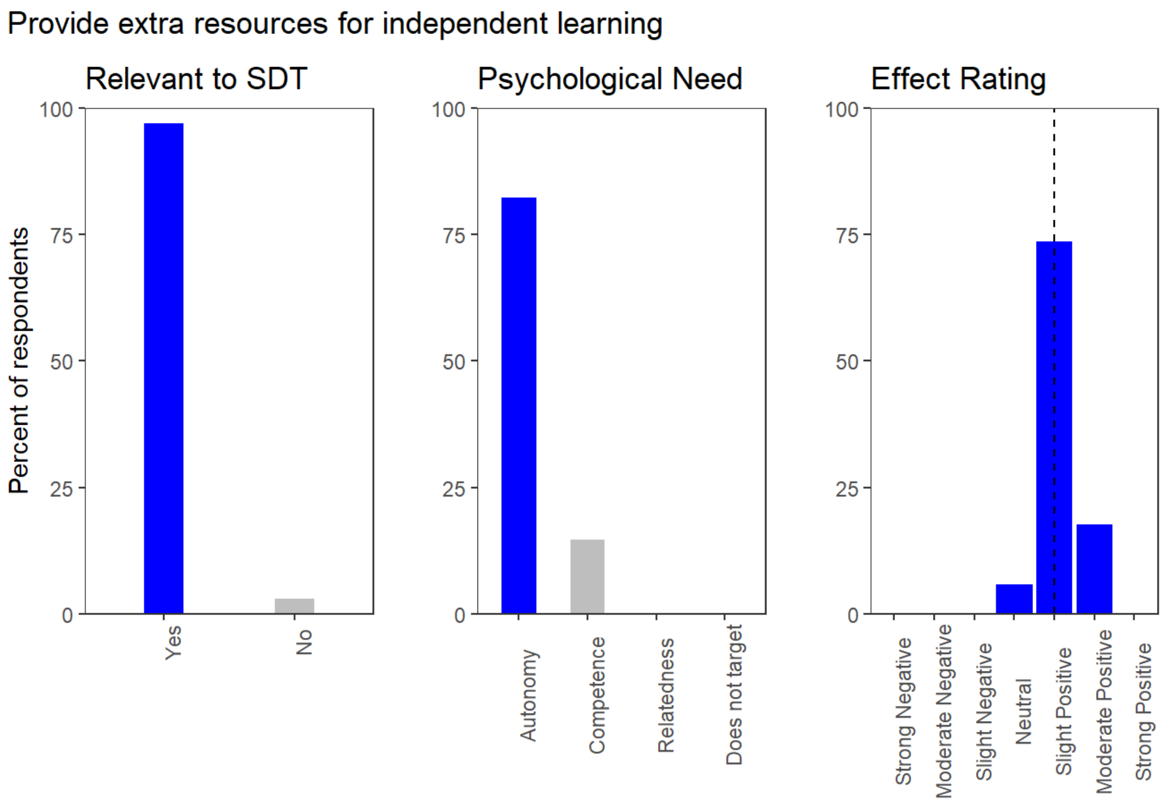
Introduce extra resources for further learning or support outside of class time

Example Behaviour:

"If you want more help, remember maths club before school tomorrow"; "here are some extra problems if you want to practice at home"

Function Description:

Allows for self-directed learning and progress outside of class time



TMB#12

Communicate in a perspective-taking way

Description:

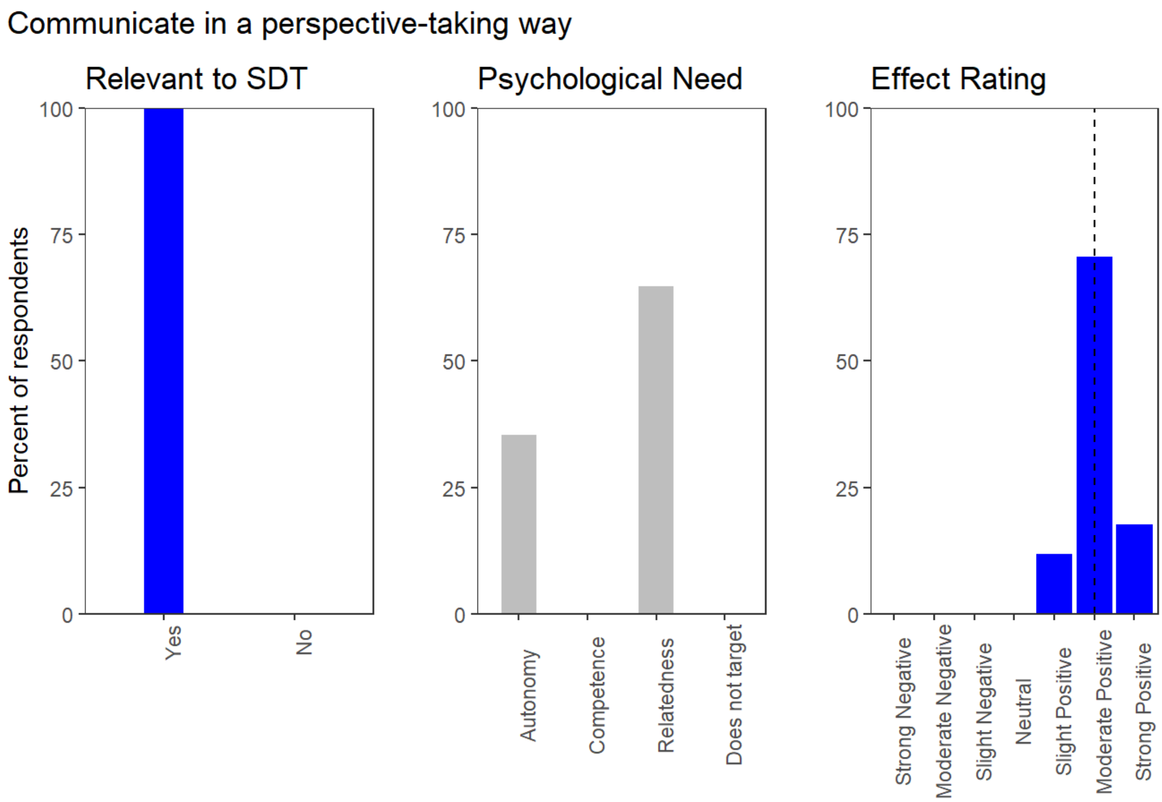
Show that you have taken a students perspective

Example Behaviour:

“Yes, you are right; this one is difficult”

Function Description:

Communicates that teacher understands the students frame of reference



TMB#13

Provoke curiosity

Description:

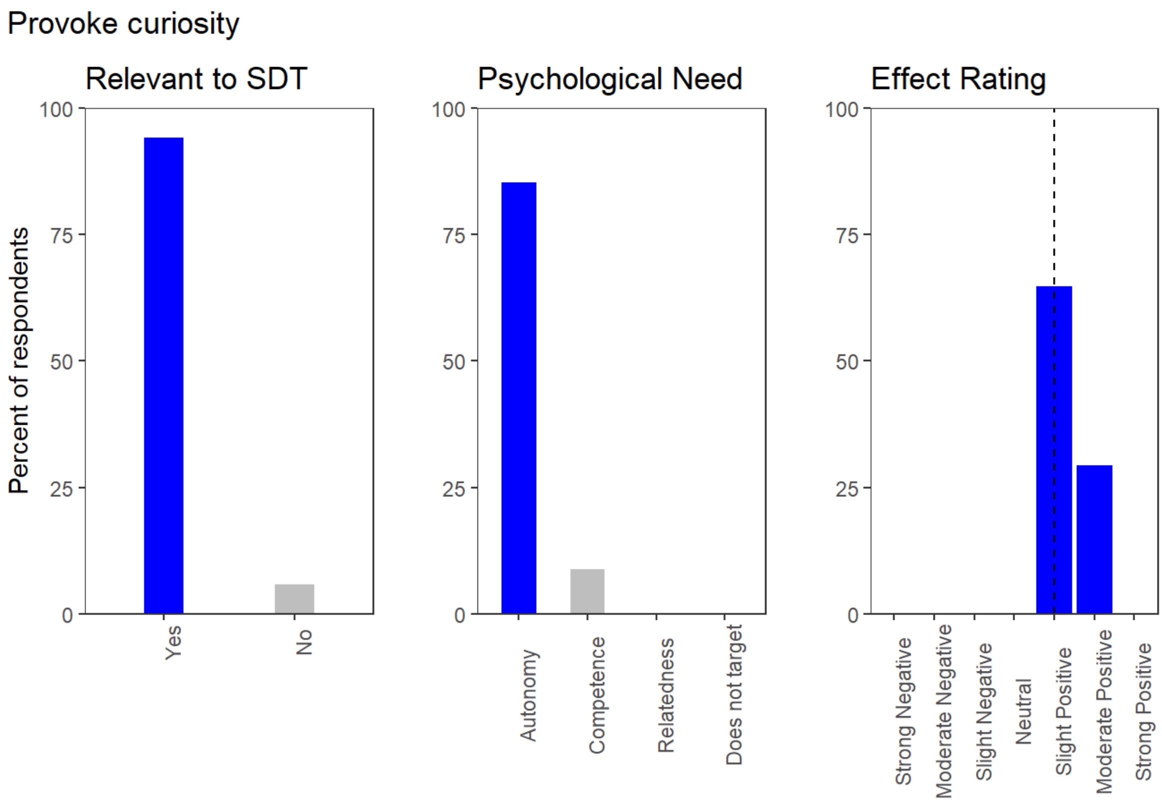
Ask a curiosity-inducing question

Example Behaviour:

"Why do we always see the same side of the moon?"

Function Description:

Piques student interest through facilitating their exploratory behaviour



TMB#14

Show understanding of the students' point of view

Description:

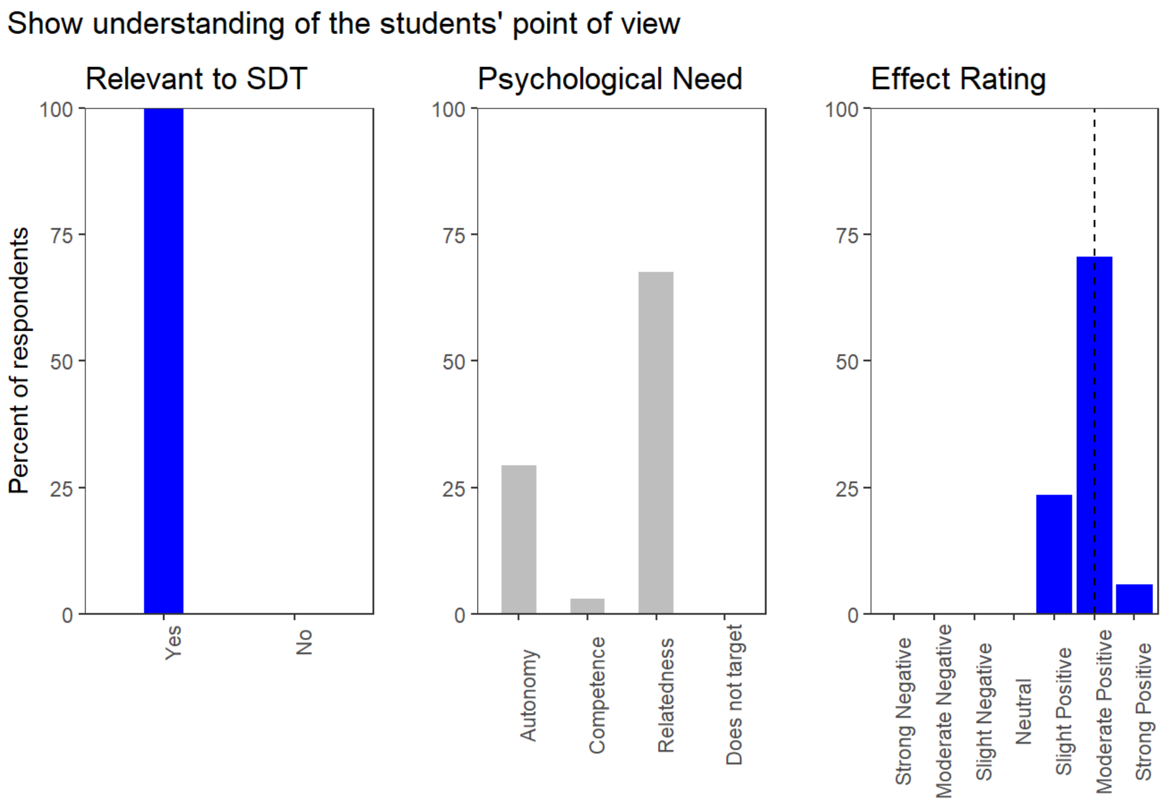
Try to understand how students see things before suggesting a new way to do things.

Example Behaviour:

"I know many of you wish I didn't assign homework today. You've said you don't like homework over the weekend."

Function Description:

Helps the student feel listened-to and understood.



TMB#15

Use pupils as positive role models

Description:

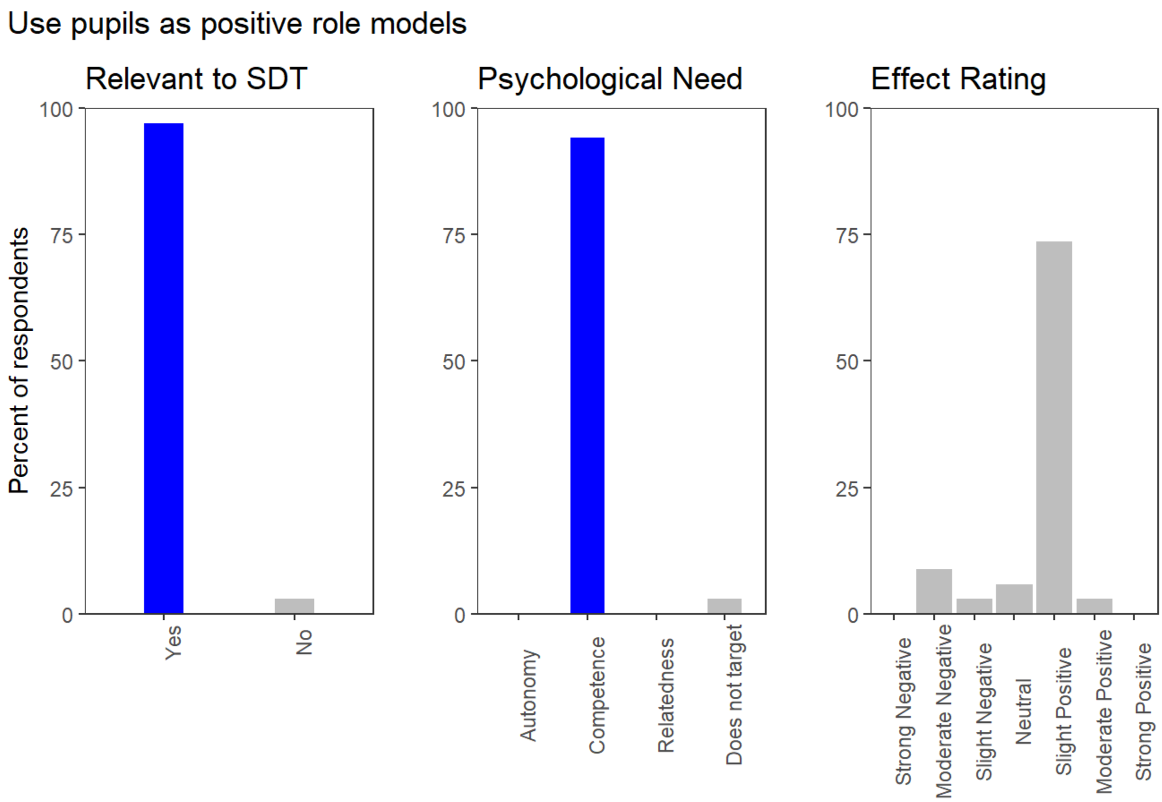
Highlight some students as examples for the rest of the class to follow

Example Behaviour:

"John, you commented on your code very well. Can we put it on the smartboard so your friends can see it?"

Function Description:

Increase self-belief through vicarious experiences of success



TMB#16

Provide rewards fairly

Description:

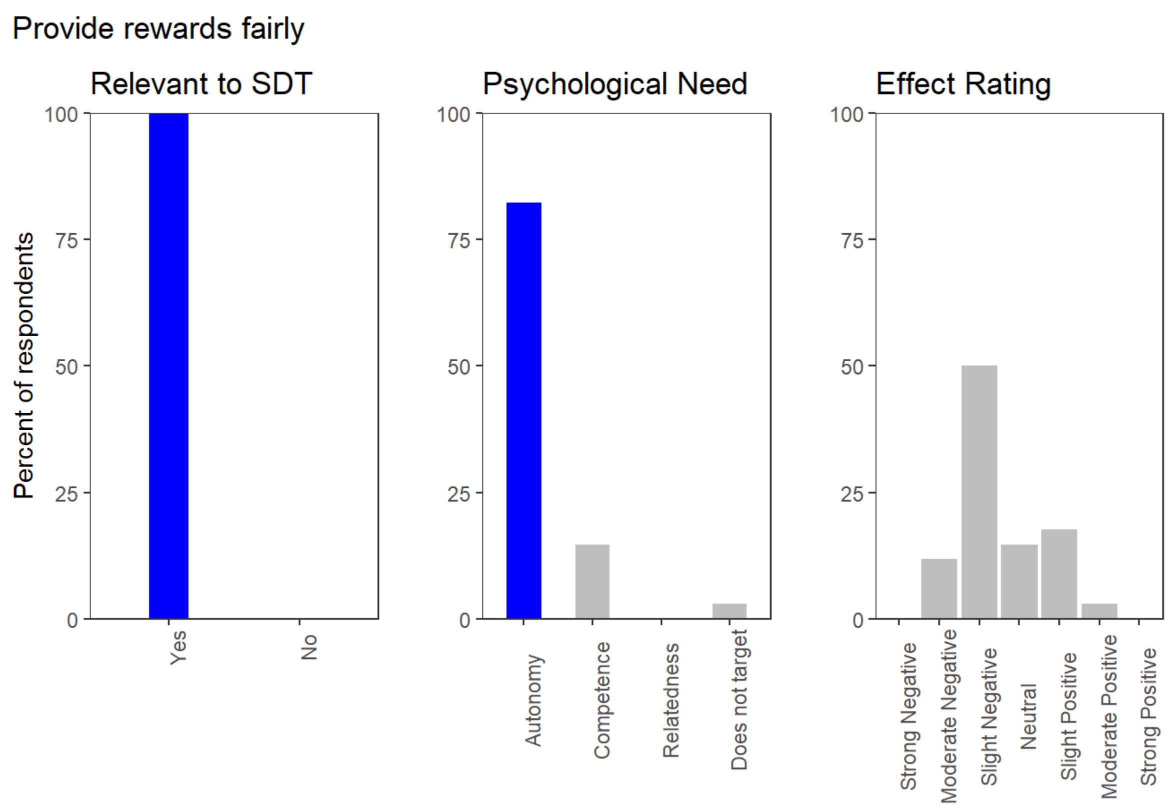
Provide rewards when the expected behaviour is observed

Example Behaviour:

"You all did your homework, so as I promised, we can watch a YouTube video today"

Function Description:

Adds external, tangible signal of which behaviours are desirable/valued by the teacher



TMB#17

Use abusive language (content)

Description:

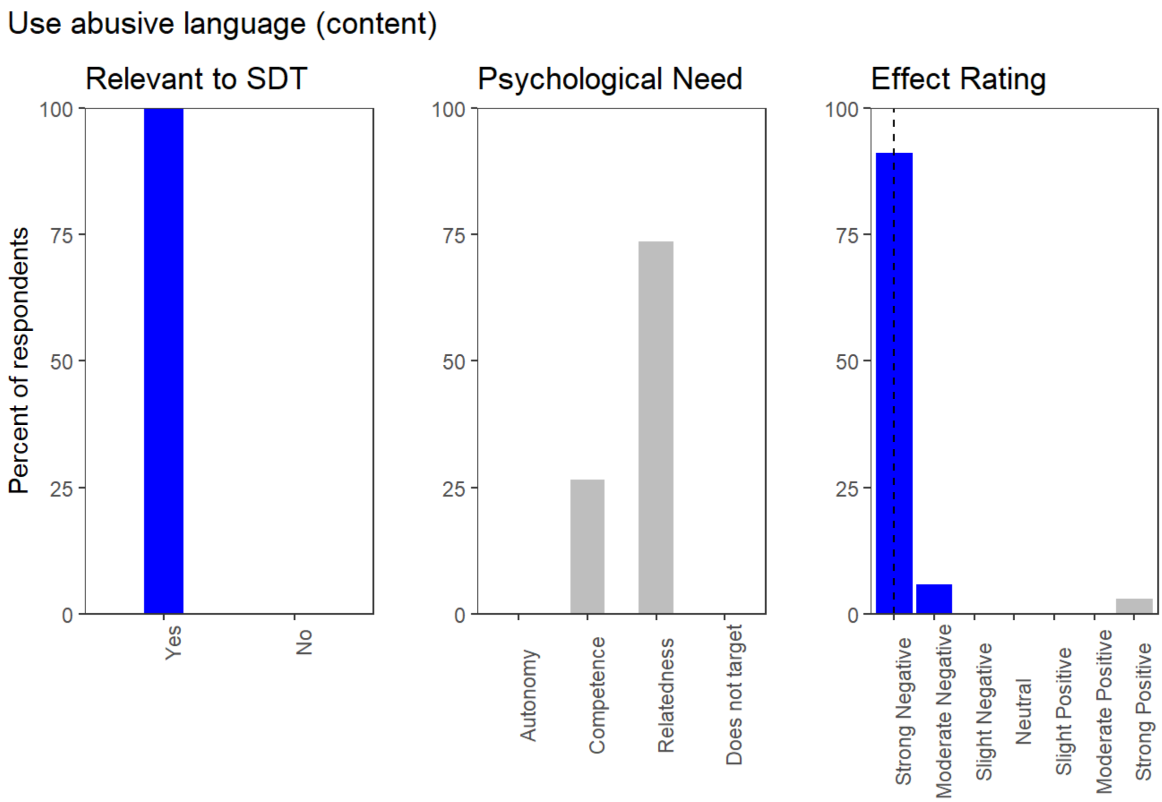
Calling students by hurtful names when they misbehave

Example Behaviour:

Calling a student "dummie" or "moron"

Function Description:

Performance mistakes and behavioural misconduct are met with competence-threatening punishment



TMB#18

Provide rewards unfairly

Description:

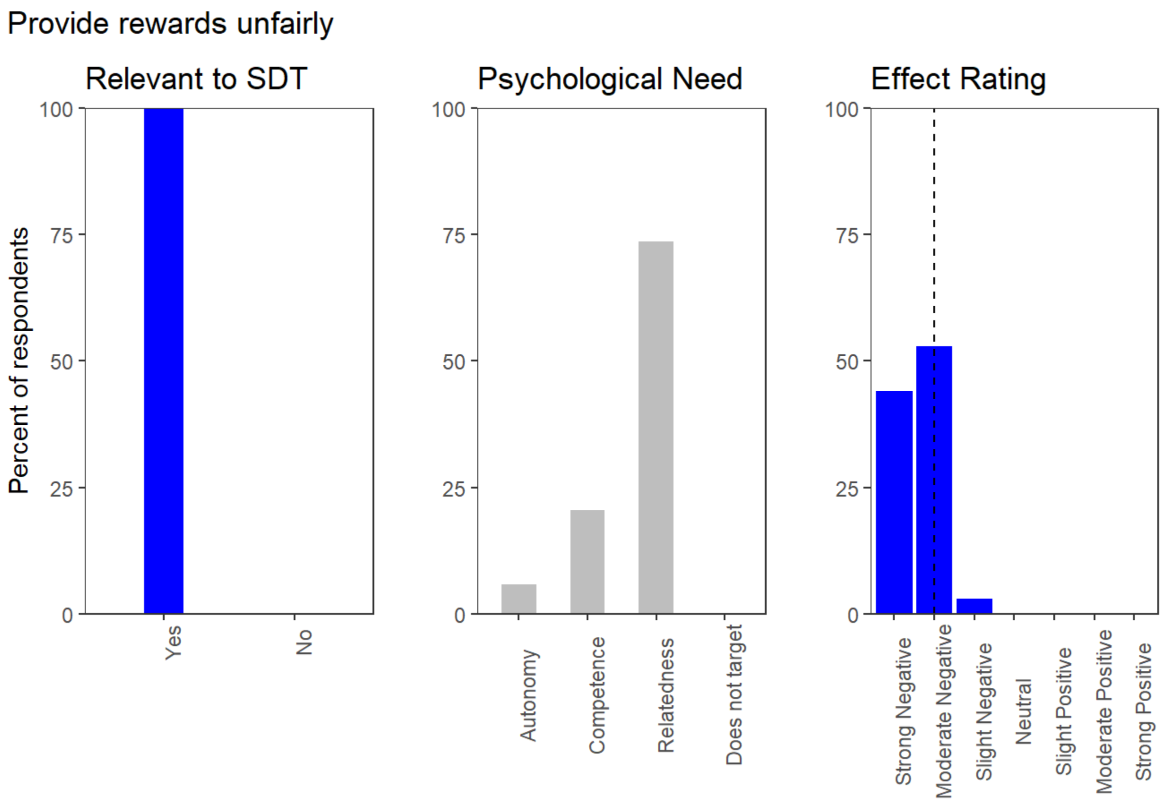
Provide rewards unfairly so students who are doing equally well, get different rewards

Example Behaviour:

Rewarding only one of three people who all completed a task

Function Description:

Students feel rewards are not predictable and teacher behaviour unjust



TMB#19

Exhibiting solutions or answers

Description:

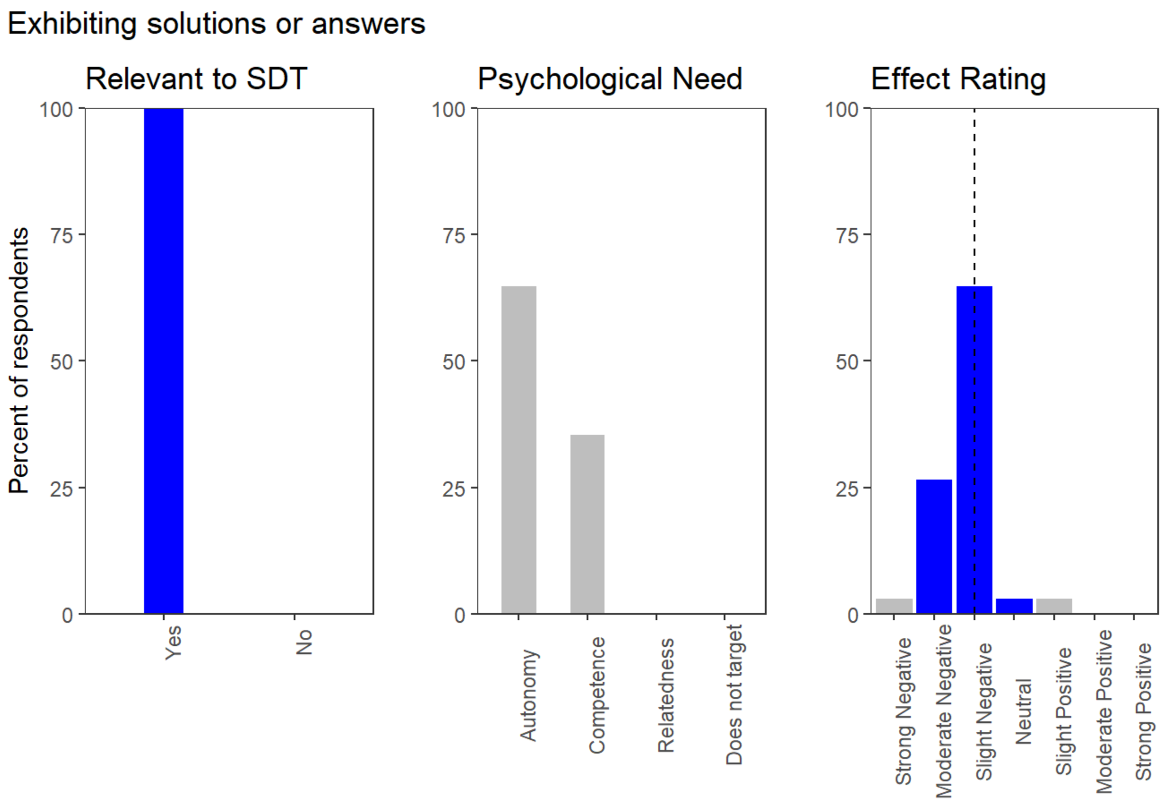
Give answers to problems instead of letting students figure it out

Example Behaviour:

"The answer is 42"

Function Description:

Stifles self-directed learning and provides external locus of causality for succes (i.e., from the teacher)



TMB#20

Praise winning via peer comparison

Description:

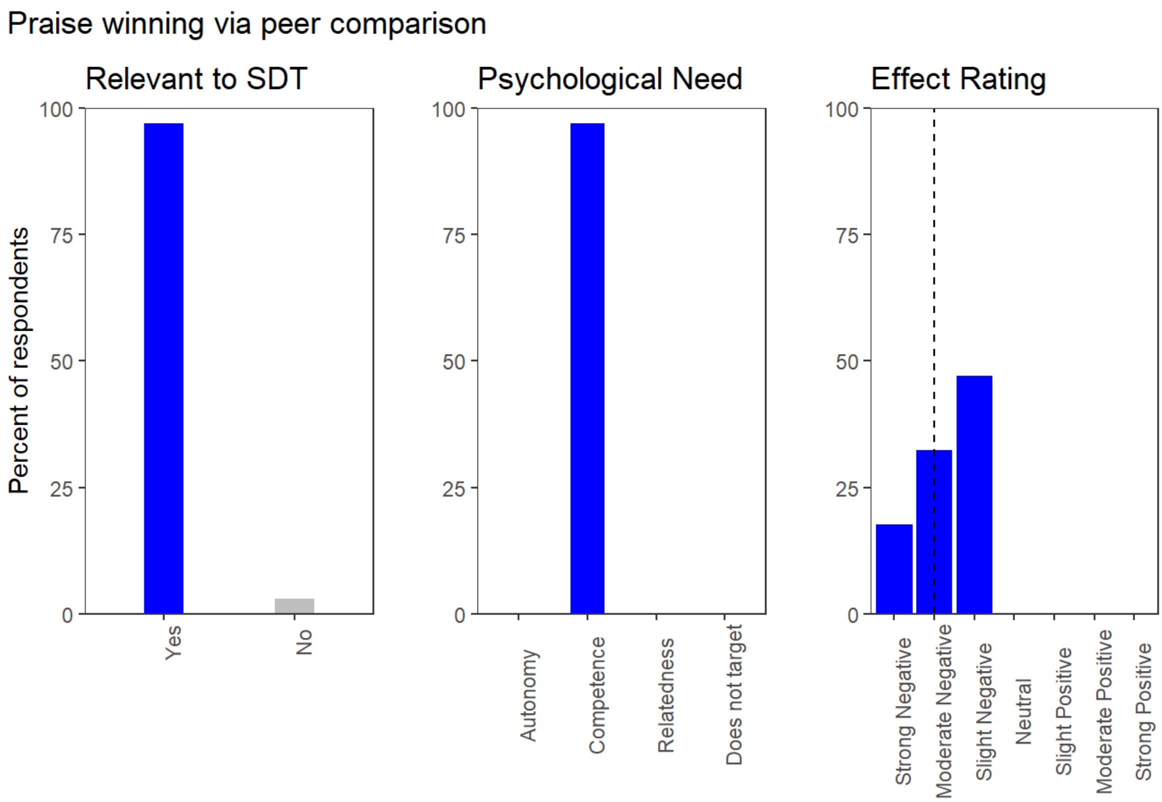
Congratulate winners so that everyone knows who did the best

Example Behaviour:

"The highest score on the exam was John"

Function Description:

Emphasises peer comparison and establishing a sense of competence, meaning few students experience success by being the best



TMB#21

Chaotic or Absent Teaching

Description:

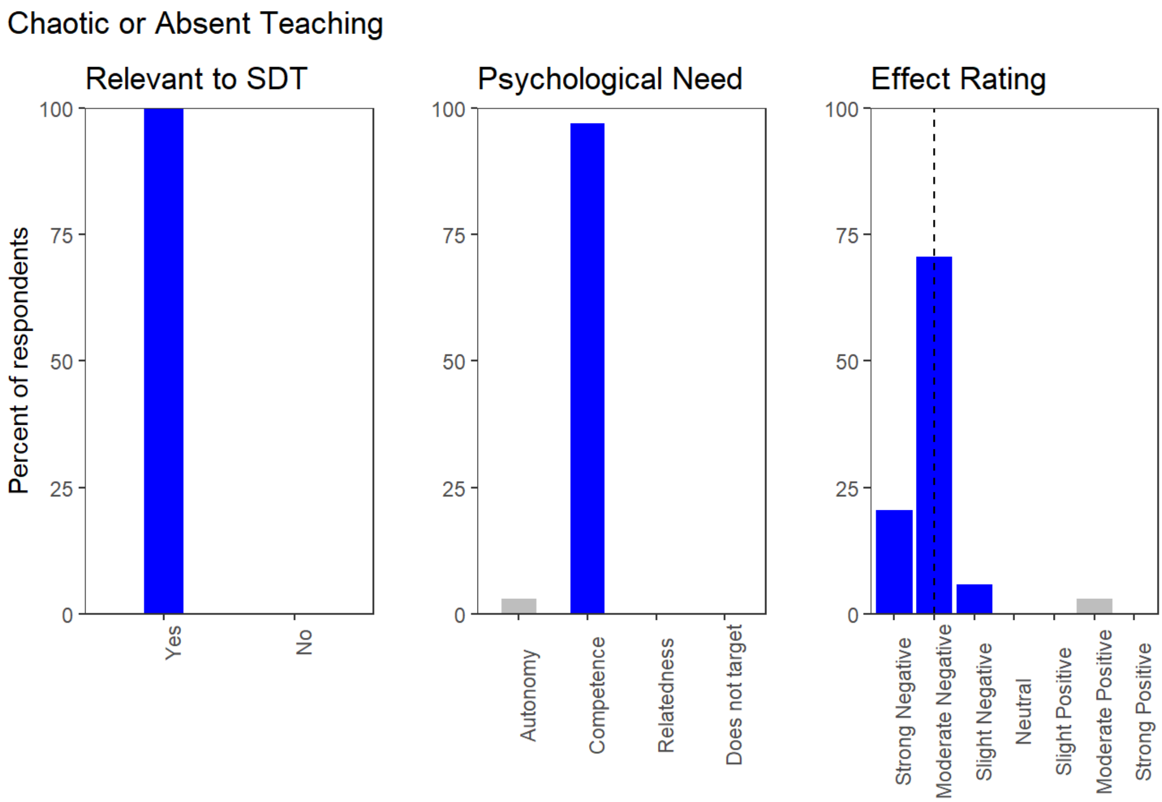
Leave students without clear instructions so the class waits or is disorganised while the teacher does something else

Example Behaviour:

Teacher leaves students waiting when arranging papers at front; Teacher gives up on providing feedback so checks his/her emails in class

Function Description:

Students do not know what they should be doing to learn and do not get any feedback or structure about how to persue goals



TMB#22

Humour

Description:

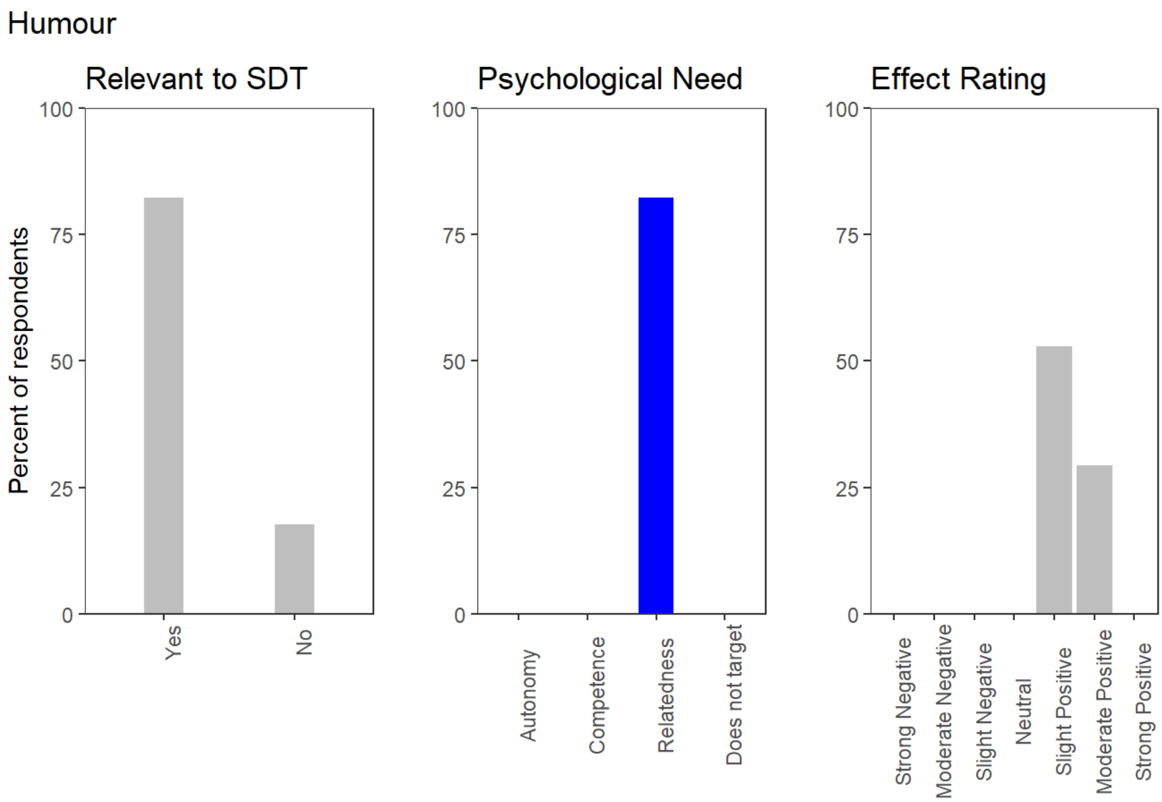
Use authentic humour so the class is fun

Example Behaviour:

"What did the triangle say to the circle? You are pointless"

Function Description:

Alleviates anxiety and reduces ego-involved goal-focus; increases warmth for teacher; stimulates interest



TMB#23

Create heterogeneous groups

Description:

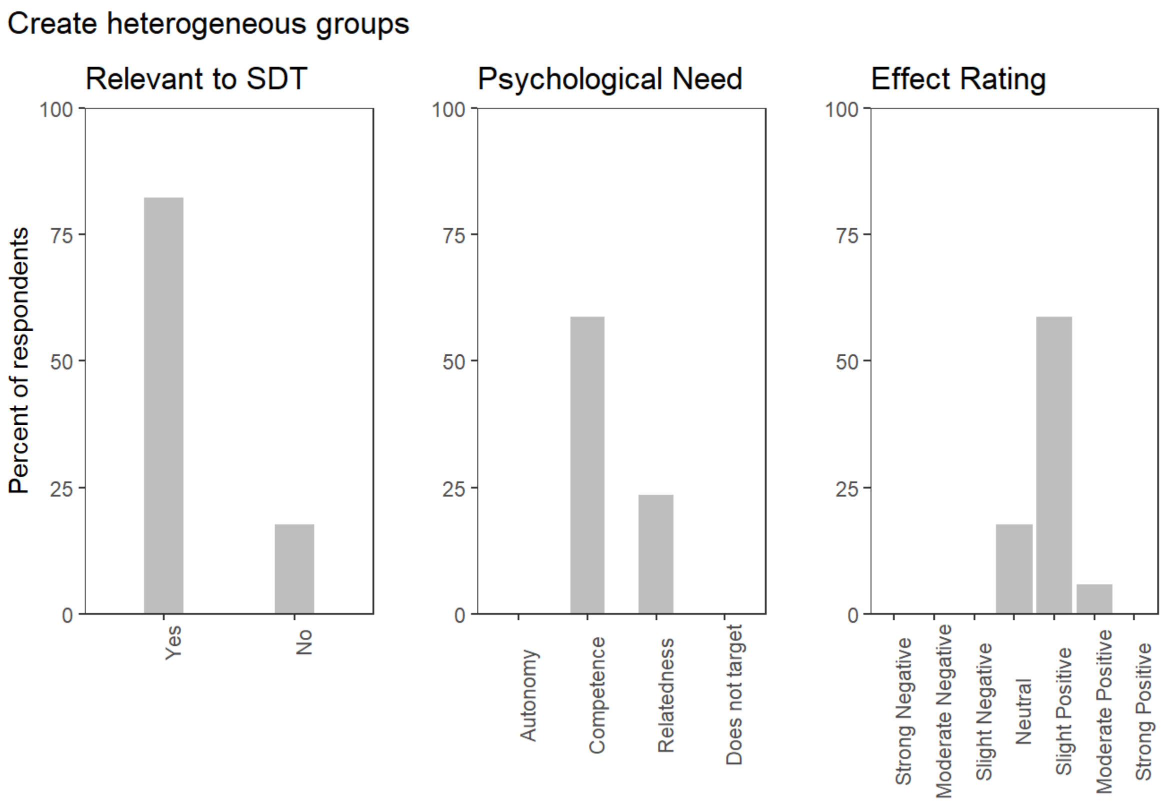
For group activities, assign students so that each group has a mix of abilities

Example Behaviour:

"Take a playing card, and find the other students with the same suit as you"

Function Description:

Removes public signalling of incompetence; increases chance of balanced frames of reference and more diverse interpersonal connections



TMB#24

Provide a variety of activities

Description:

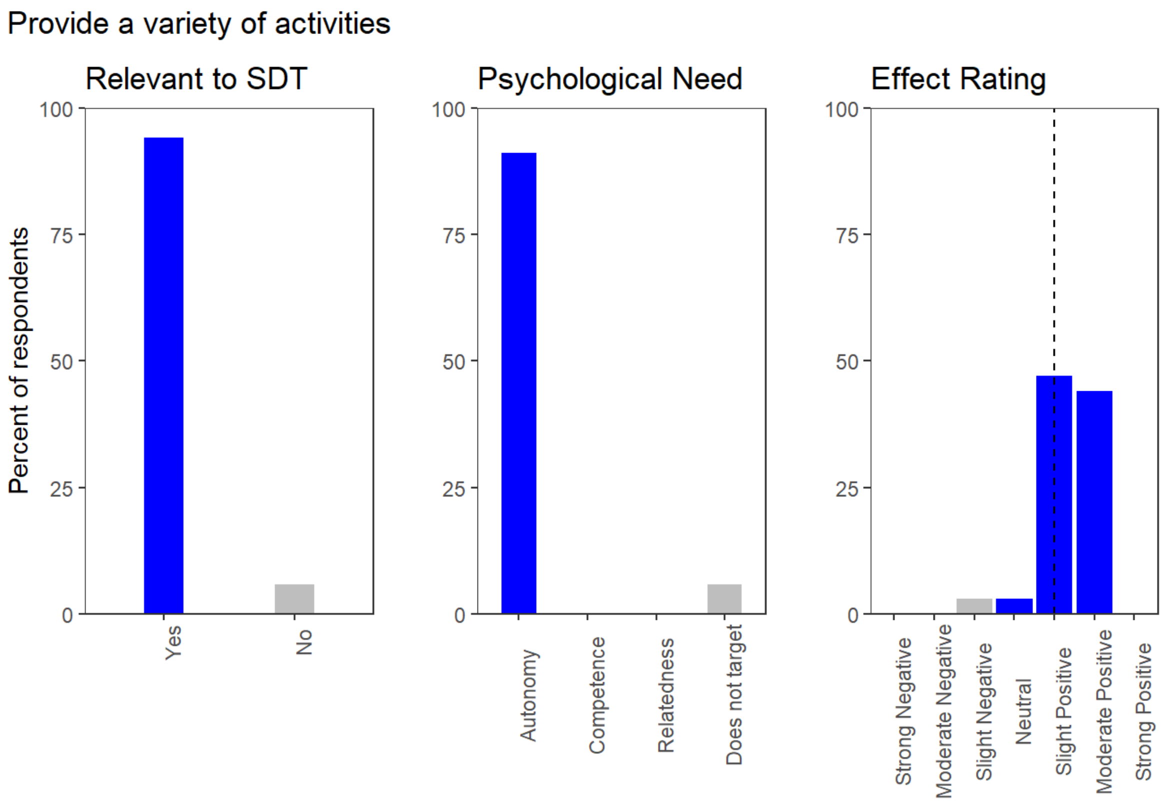
Provide a variety of activities in a way that keeps things interesting

Example Behaviour:

Teacher regularly changes the format of the class (debates one lesson, worksheets the next), and presents content in dynamic ways (teaches US History using Hamilton)

Function Description:

Reduces boredom



TMB#25

Apply fair punishments

Description:

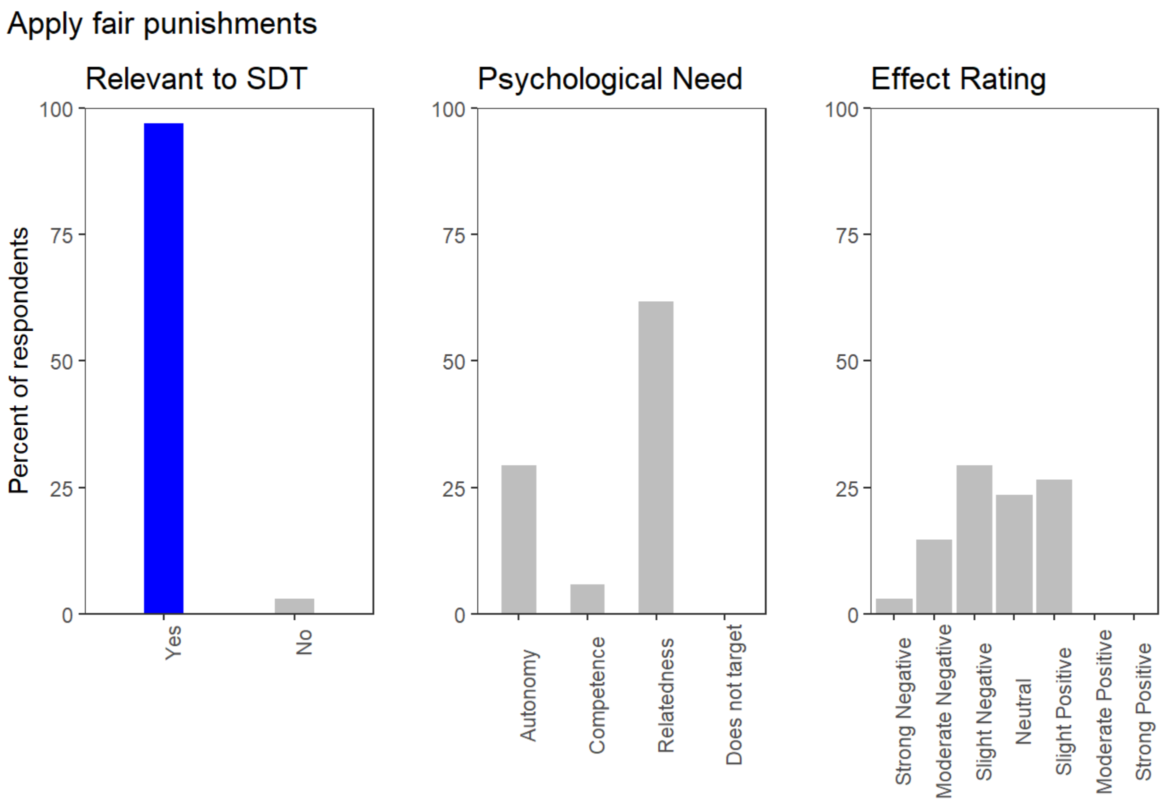
Provide punishments fairly so students who misbehave are treated equally

Example Behaviour:

Sending both of two students out of class when they misbehave or break a rule

Function Description:

Ensures misbehaviour is consistently and reliably met with external contingencies



TMB#26

Set up activities that exclude some students

Description:

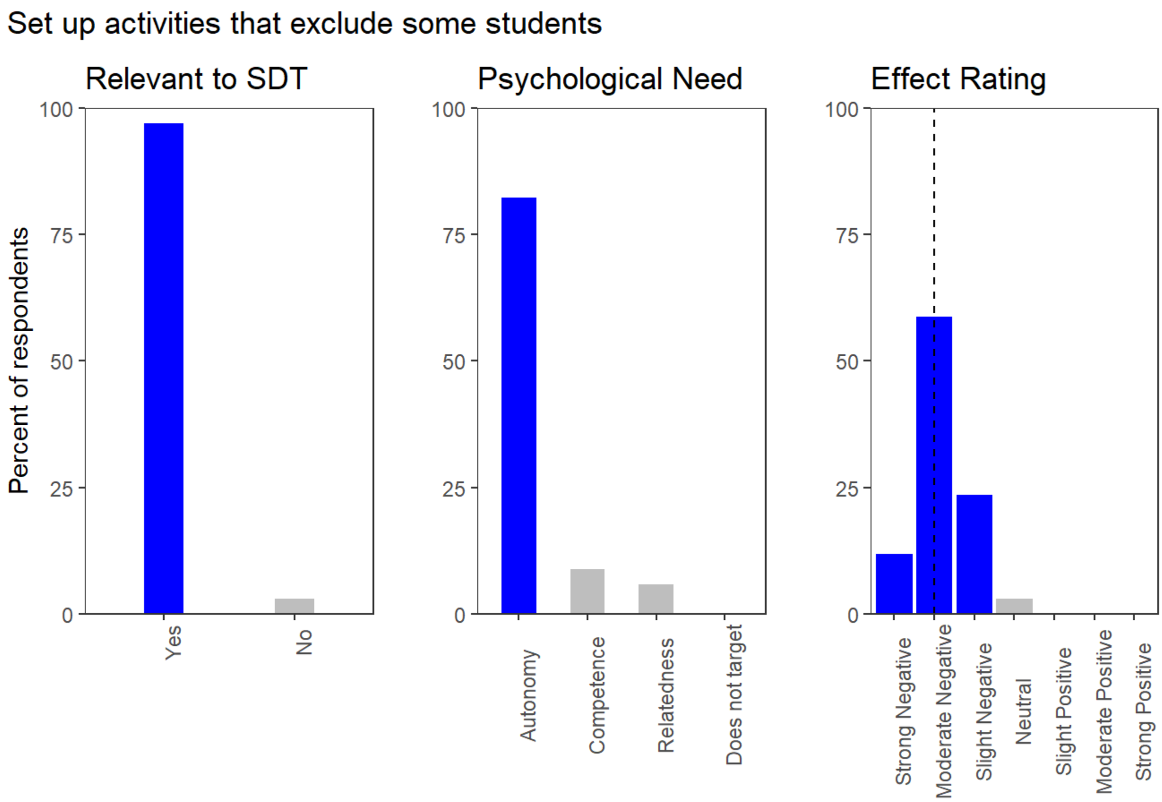
Set up activities so there are times where some students are not doing anything

Example Behaviour:

"if you have finished the questions, just sit quietly until everyone else is finished"

Function Description:

Students do not have opportunities to engage even if they want to



TMB#27

Teaching students to set intrinsic life goals for learning

Description:

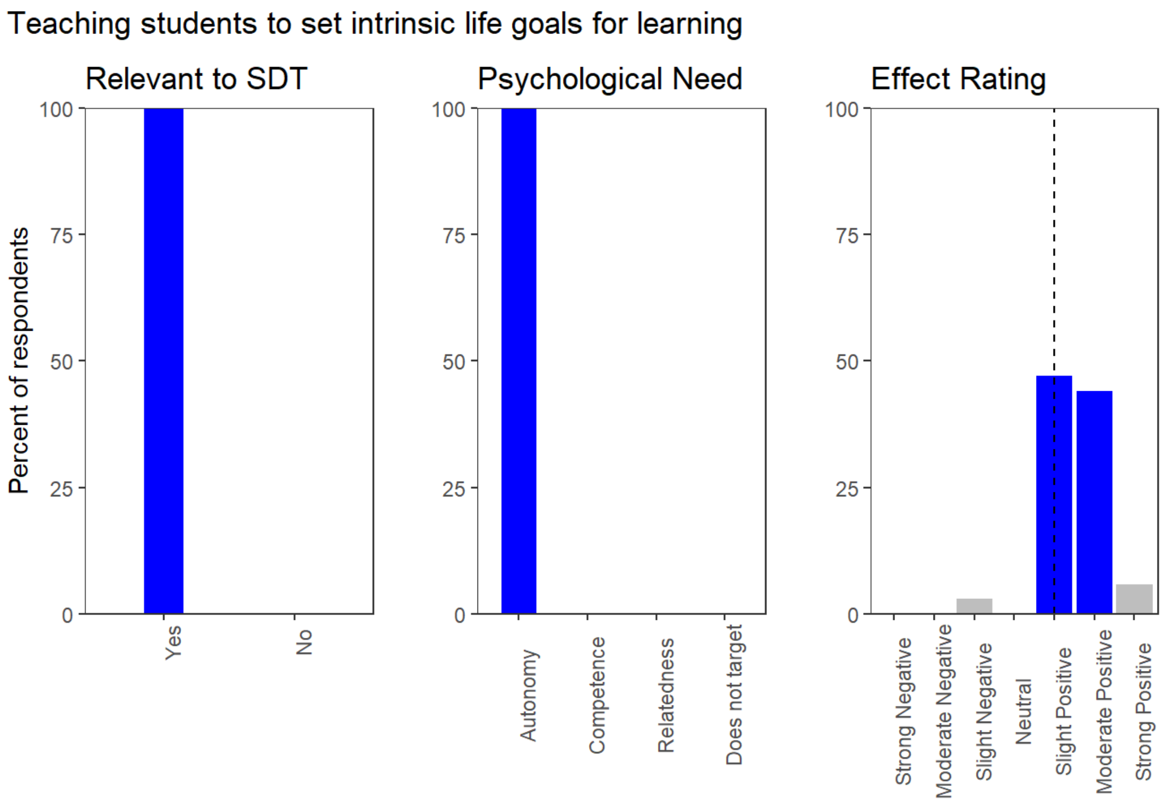
Help students link learning to other intrinsic life goals, like helping others, being healthy, embracing challenges, or improving the world.

Example Behaviour:

"Reading helps me to gain knowledge about life" or "I want to use my reading skills to read to little kids".

Function Description:

Students will try to understand the lessons more, become better at doing the activities, so that students can help others someday, or discover something interesting.



TMB#28

Creating meaning through using parables, stories, analogies, or metaphors

Description:

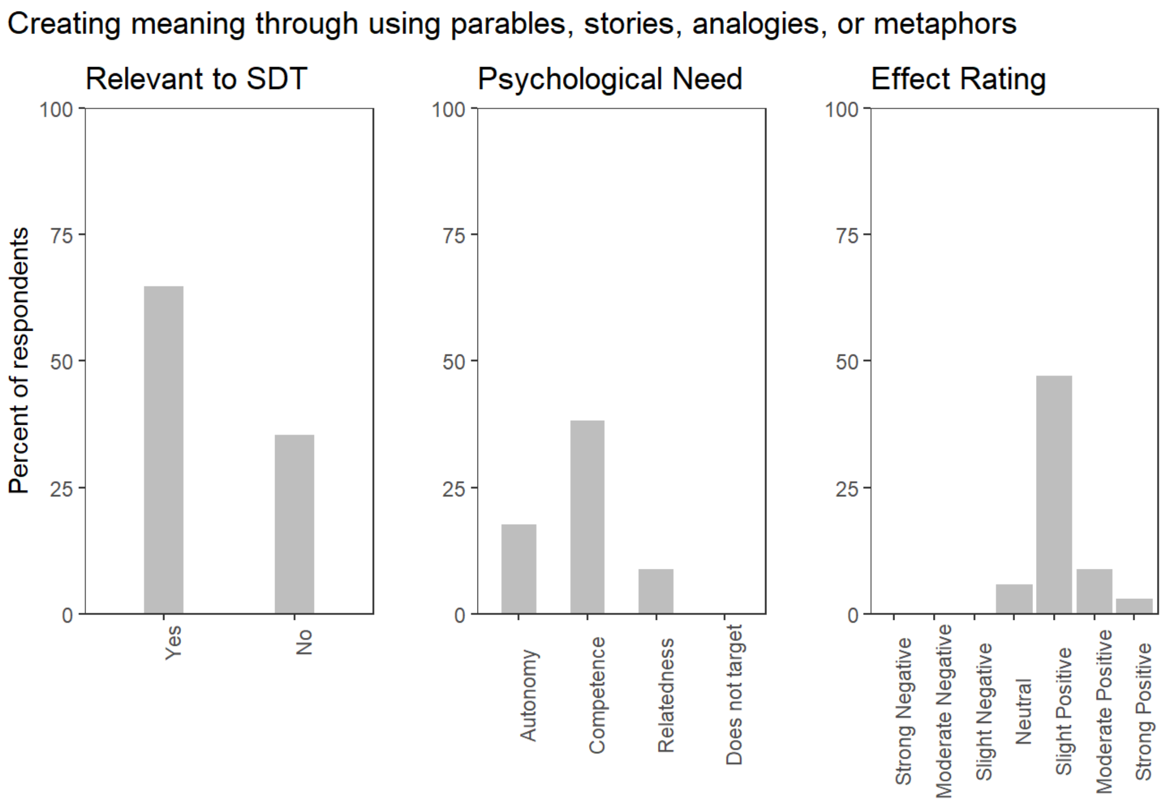
Use parables, stories, analogies, or metaphors to help students to connect abstract constructs into concrete examples

Example Behaviour:

"mistakes are stepping stones, not stumbling blocks"; "treating others well is like sowing good seeds, eventually you'll reap a good harvest"

Function Description:

Promote empathy through narratives that students can connect to, and provide examples that students could follow.



TMB#29

Set goals on behalf of the class

Description:

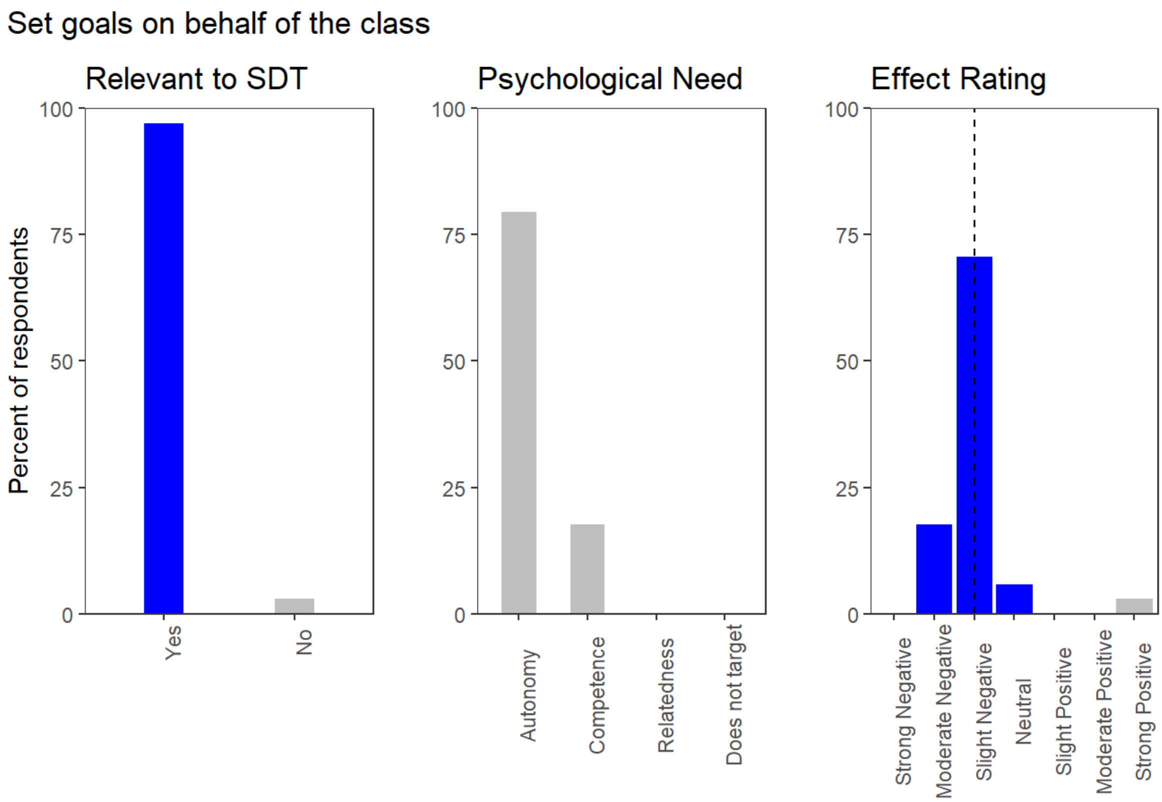
Teachers setting goals for students rather than letting them decide their own goals

Example Behaviour:

"Complete all questions on page 4 by the start of class tomorrow. I want you all to get at least 9 out of 10."

Function Description:

Provides the same level of challenge for the whole class.



TMB#30

Provide incorrect modelling examples

Description:

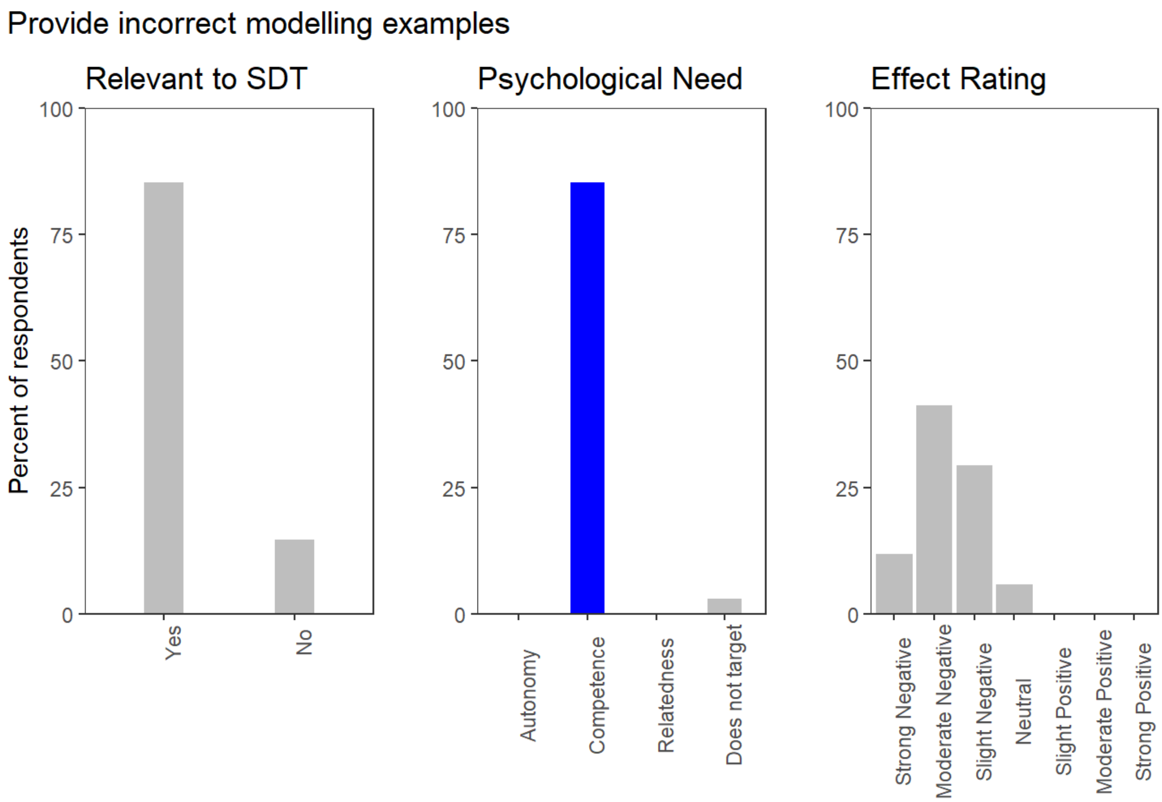
The teacher models in a way that is unlikely to produce the desired outcome without correcting behaviour.

Example Behaviour:

"Helium is the first element of the periodic table"

Function Description:

Undermines progress because students are likely to reproduce the undesirable approach.



TMB#31

Group students with similar interests

Description:

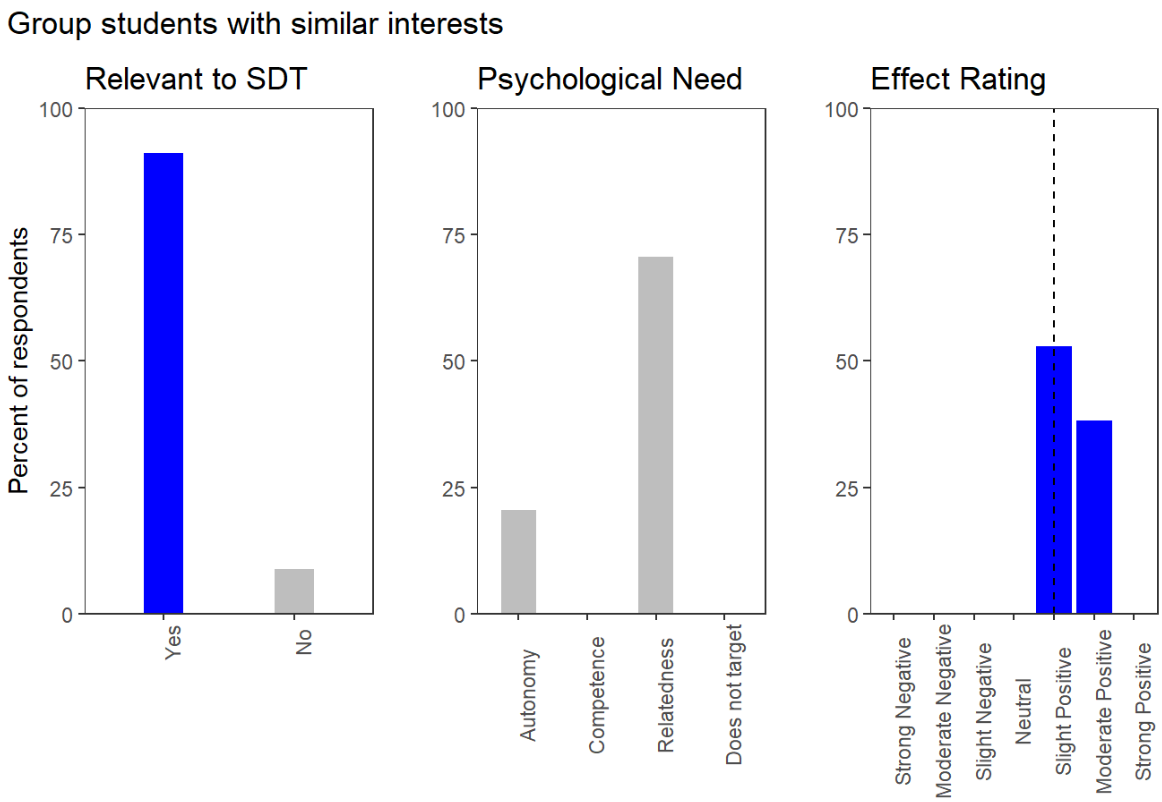
Create groups in the class where students with similar values or interests can work together on problems

Example Behaviour:

When studying geography, grouping musical students to look at a country's music, the sporty students to look at the country's sports, and other students to look at the country's key historical events.

Function Description:

Allows students to work on tasks—and with people—that match their interests and values.



TMB#32

Regular communication with parents

Description:

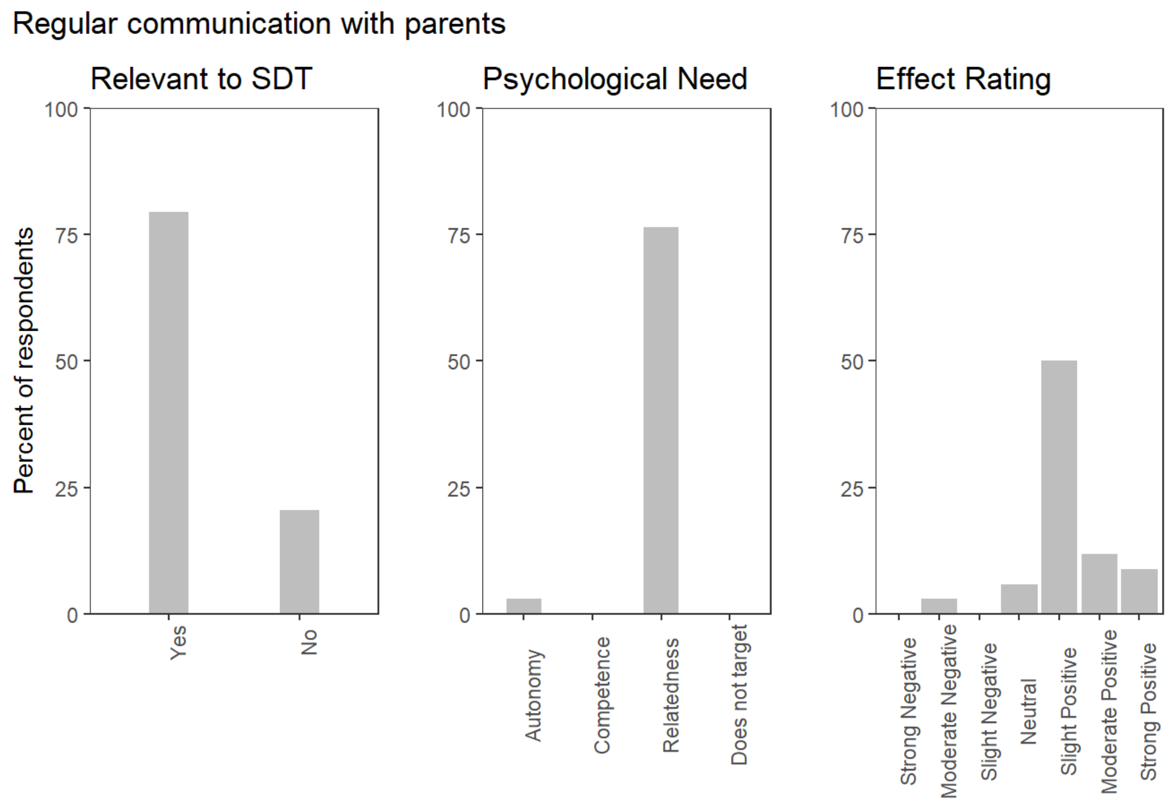
Teachers engaging in regular contact (e.g., phone, email, text) with parents about the activities in the class,

Example Behaviour:

The teacher calls a parent when she notices that a student has been particularly disengaged and unenthusiastic to talk and when the children improved, was particularly engaged and/or creative.

Function Description:

If contact is not transactional, it may support connections between the students' home and school life, identifying ways that key people in each domain (e.g., parents, teachers) can support each-other, building trusting relationships



TMB#33

Allow for student input or choice

Description:

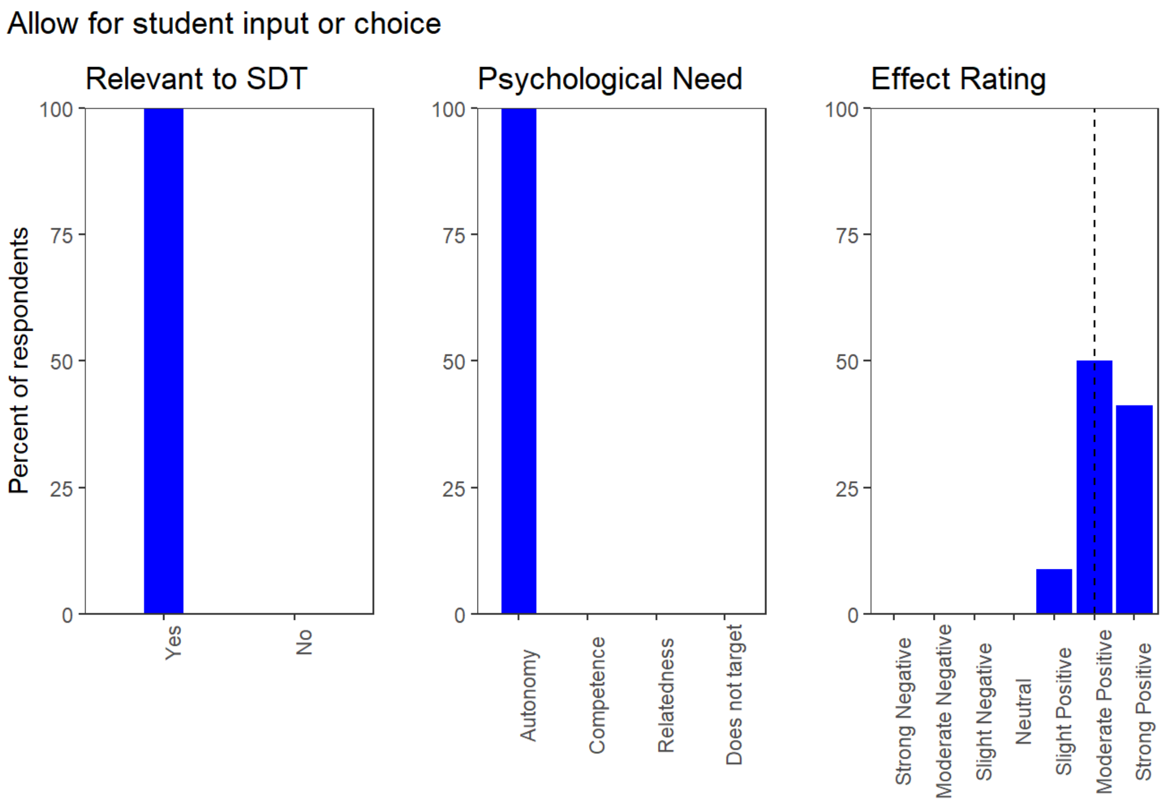
Create opportunities for students to meaningfully direct the activities they do in class

Example Behaviour:

"Feel free to work with a friend or do it by yourself"

Function Description:

Allows students to choose tasks that align with their priorities and capabilities; supports the ownership of the behaviour



TMB#34

Provide conditional positive regard

Description:

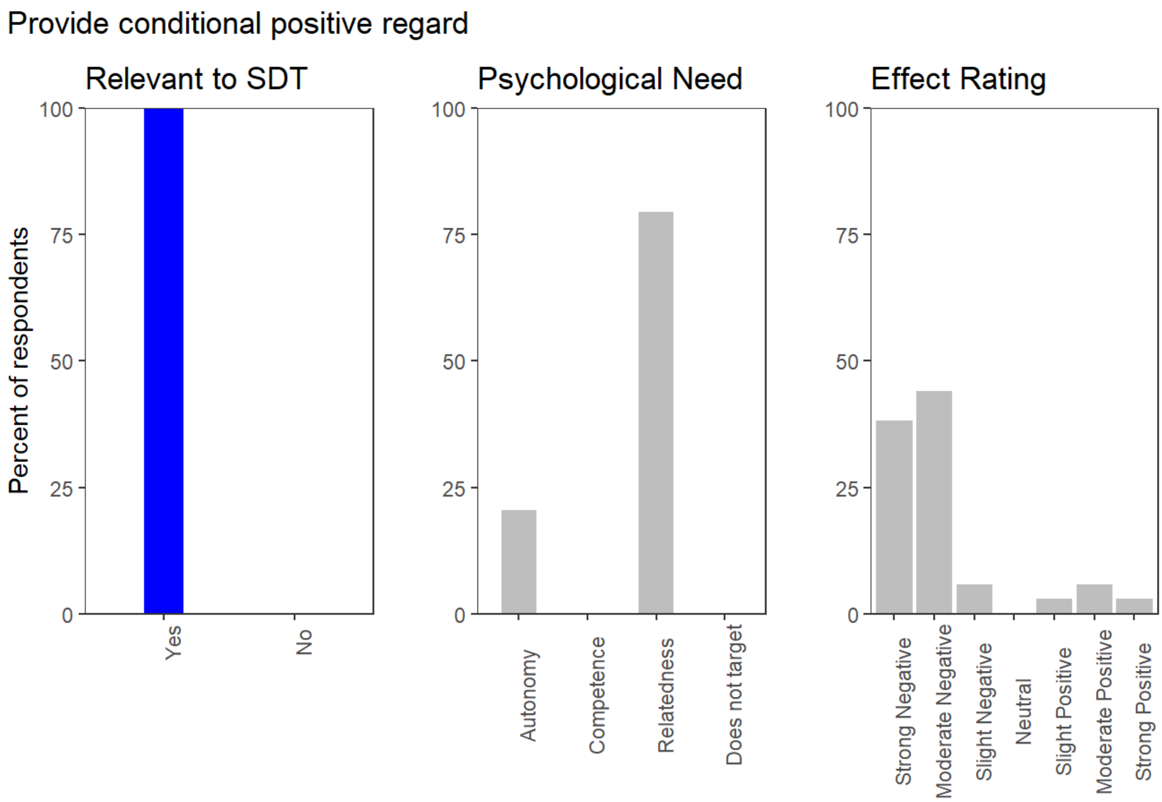
Withdrawal warmth from a student in response to poor behaviour; provide warmth and acceptance only when teacher's expectations are met

Example Behaviour:

"Good job! You did it the way I asked you!"

Function Description:

Demonstrate that attention and warmth are contingent upon meeting the teachers expectations of good student behaviour



TMB#35

Grouping students on the basis of ability

Description:

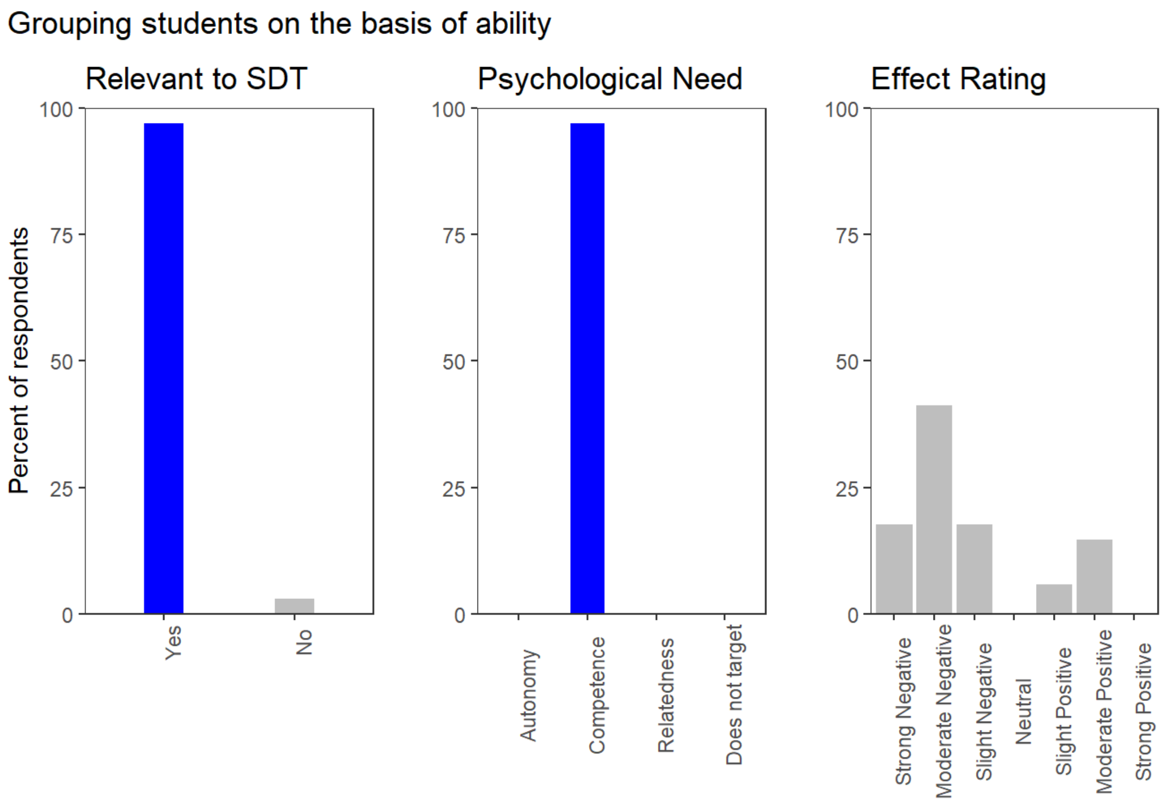
Grouping is done publicly and students are put in groups based on their ability so that there are "top" and "bottom" groups

Example Behaviour:

"In the green group, you are a beginner in handstand. In the blue group, you are a little bit experienced in doing a handstand, and in the black group, you are able to make a very good handstand."; "If you got more than 7/10, join this group. Less than 7: in this group. If you did not do your homework, you are at the back"

Function Description:

Increases public signalling of student competence, and means students are comparing themselves to others of similar abilities



TMB#36

Ask students about their experience of lessons

Description:

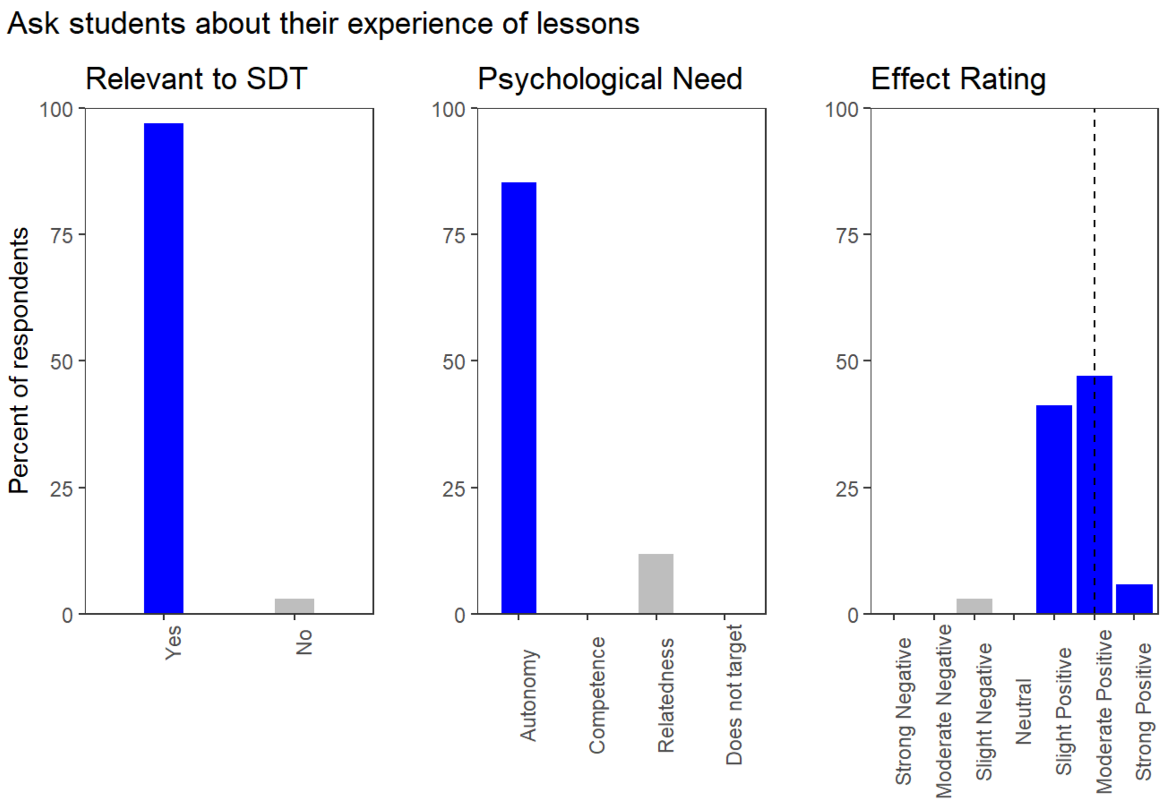
Ask students for feedback about how classes are going; could apply to either the content of lessons or the process/learning design

Example Behaviour:

"On these sheets, please write down what you liked about today's lesson, what you didn't like, and what was most unclear. Remember it's anonymous."

Function Description:

Gives students a safe opportunity to shape the way classes are run, so lessons can better cater to their needs and interests


Ask students about their experience of lessons
Relevant to SDT
Psychological Need
Effect Rating
Percent of respondents
100
75
50
25
0
Yes
No
Autonomy
Competence
Relatedness
Does not target
Strong Negative
Moderate Negative
Slight Negative
Neutral
Slight Positive
Moderate Positive
Strong Positive

```
```

Appendix C.4

Results for Removing/Retaining Behaviours

TMBs	Go	Stay	Consensus
Set goals where students compete against each-other	12	19	Stay
Provide transparent structure	16	15	Go
Responding to Queries	16	15	Go
Offer rewards	25	6	Go
Ask questions to expand understanding	14	17	Stay
Provide regular praise	26	5	Go
Provide rewards fairly	25	6	Go
Adopting student initiatives	22	9	Go
Ask students about their experience of lessons	15	16	Stay
Communicate in a perspective-taking way	26	5	Go
Create heterogeneous groups	27	4	Go
Set goals on behalf of the class	23	8	Go
Provide frequent constructive criticism	30	1	Go
Provide feedback aimed at improvement or effort	14	17	Stay
Provide students with personal attention	17	14	Go
Show interest in outside school activities	20	11	Go
Provide feedback in private	12	19	Stay
Praise a student in public	18	13	Go
Outlining Punishment Contingencies	20	11	Go
Ask controlling questions	22	9	Go
Uttering directives or commands	25	6	Go

Praise a student's fixed qualities	18	13	Go
Total Consensus'	17	5	-

Appendix C.5

All the deleted behaviours (TMBs) in rounds 1, 2, and 3

Reached Consensus?	**Behaviour**	**Description**	**Example Behaviour**	**Function Description**
Deleted (overlapping)	Uttering directives or commands	Command students to do things without providing rationales.	"Do it like this."	Imposes expectations on students without aligning it to any reasons.
Deleted (overlapping)	Show interest in outside school activities	Express interest in students activities outside of school	"What sports did you do on the weekend?"	Shows student is recognised and valued as a person, not just a student
Deleted (overlapping)	Provide students with personal attention	Pay close attention to students' progress and wellbeing while working in class	Moving between tables while students collaborate on a difficult problem	Makes students feel that their work is important and their efforts are noticed
No	Active Learning	Set up activities where all students are engaged in a learning activity	Complete this worksheet individually to figure out how heavy the Sydney Harbour Bridge is; "Try to make a sentence using as few of these phonemes as possible"	Allows each student hands-on practice with an activity designed to progress development of a skill
No	Discuss class values	Collaboratively establish the values important to display in the class, or remind students of the collaboratively derived values	We all thought working hard was important, so even though many find this task difficult, see if you can push through to the end.	Connects the activities that take place in class with values that the student cares about
No	Teacher enthusiasm	Present content enthusiastically to make things fun and interesting	Now I think this next part of the lesson is really interesting!	Models the attitude and energy that the teacher would like the students to demonstrate; shows interest in the material.

No	Offering hints	Give hints to help students along without giving them the "right answer"	“It might be easier to start with this formula.”	Supports the student’s own learning processes. Allows students to maintain an internal locus of causality during learning.
No	Provide extra resources for independent learning	Introduce extra resources for further learning or support outside of class time	If you want more help, remember maths club before school tomorrow; "here are some extra problems if you want to practice at home"	Allows for self-directed learning and progress outside of class time
Deleted (overlapping)	Provide transparent structure	Provide an overview of what we are going to do in the lesson	In todays class, we are working on ratios in three ways...	Provides a plan for students to follow so they know how things are going to be organised
Deleted (overlapping)	Responding to Queries	Answer student questions fully and carefully	Not quite, that is the formula for Sin not Cos.	Clarifies path toward goal achievement.
Deleted (overlapping)	Acknowledge student negative feelings	Acknowledge students negative feelings	I noticed you are looking frustrated.	Validates emotions as understandable, normal, and expected
No	Provoke curiosity	Ask a curiosity-inducing question	Why do we always see the same side of the moon?	Piques student interest through facilitating their exploratory behaviour
No	Show understanding of the students' point of view	Try to understand how students see things before suggesting a new way to do things.	I know many of you wish I didn't assign homework today. You've said you don't like homework over the weekend.	Helps the student feel listened-to and understood.
No	Use pupils as positive role models	Highlight some students as examples for the rest of the class to follow	John, you commented on your code very well. Can we put it on the smartboard so your friends can see it?	Increase self-belief through vicarious experiences of success
Deleted (overlapping)	Ask controlling questions	Provide commands that are phrased as	Can you do it like I showed you?” Communicates	Communicates disapproval for the students current

		rhetorical questions	disapproval for the students current behaviour; reinforces need for obedience; promotes students' defensive self-justifications	behaviour; reinforces need for obedience; promotes students' defensive self-justifications
Deleted (overlapping)	Outlining Punishment Contingencies	Declaring (but not yet enforcing) if-then extrinsic punishments—contingencies that are not inherent to the task and are provided in an effort to extinguish a behaviour	If you two speak one more time, I will send you out	Imposes an extrinsic reason for student behaviour.
No	Use abusive language (content)	Calling students by hurtful names when they misbehave	Calling a student "dummie" or "moron"	Performance mistakes and behavioural misconduct are met with competence-threatening punishment
No	Provide rewards unfairly	Provide rewards unfairly so students who are doing equally well, get different rewards	Rewarding only one of three people who all completed a task	Students feel rewards are not predictable and teacher behaviour unjust
No	Exhibiting solutions or answers	Give answers to problems instead of letting students figure it out	The answer is 42	Stifles self-directed learning and provides external locus of causality for succes (i.e., from the teacher)
No	Praise winning via peer comparison	Congratulate winners so that everyone knows who did the best	The highest score on the exam was John	Emphasises peer comparison and establishing a sense of competence, meaning few students experience success by being the best

No	Chaotic or Absent Teaching	Leave students without clear instructions so the class waits or is disorganised while the teacher does something else	Teacher leaves students waiting when arranging papers at front; Teacher gives up on providing feedback so checks his/her emails in class	Students do not know what they should be doing to learn and do not get any feedback or structure about how to persue goals
No	Provide a variety of activities	Provide a variety of activities in a way that keeps things interesting	Teacher regularly changes the format of the class (debates one lesson, worksheets the next), and presents content in dynamic ways (teaches US History using Hamilton)	Reduces boredom
Deleted (overlapping)	Ask questions to check knowledge	Ask the students clarification questions that check what students know	What is the B in BODMAS?	Fosters common understanding of goal-directed behaviours
Deleted (redundant)	Prefer open-ended questions over closed questions	Ask questions that require many words to answer	Ask questions starting with "why", "how", or "what" rather than "do", "is", or "are"	Facilitates student self-expression and deeper thinking
No	Apply fair punishments	Provide punishments fairly so students who misbehave are treated equally	Sending both of two students out of class when they misbehave or break a rule	Ensures misbehaviour is consistently and reliably met with external contingencies
No	Set up activities that exclude some students	Set up activities so there are times where some students are not doing anything	if you have finished the questions, just sit quietly until everyone else is finished	Students do not have opportunities to engage even if they want to
No	Teaching students to set intrinsic life goals for learning	Help students link learning to other intrinsic life goals, like helping others, being healthy,	Reading helps me to gain knowledge about life or "I want to use my reading skills to read to little kids".	Students will try to understand the lessons more, become better at doing the activities, so that students can help

		embracing challenges, or improving the world.		others someday, or discover something interesting.
Deleted (overlapping)	Adopting student initiatives	Take student suggestions into learning activities when they arise	That's a great idea. We can do that activity in this session.	Affirms student initiative and self-management of learning.
No	Group students with similar interests	Create groups in the class where students with similar values or interests can work together on problems	When studying geography, grouping musical students to look at a country's music, the sporty students to look at the country's sports, and other students to look at the country's key historical events.	Allows students to work on tasks—and with people—that match their interests and values.
Deleted (overlapping)	Unfair use of praise	Praises students unfairly or unequally; shows favourites	Complementing only one of three students who completed a problem a creative way	Makes students feel like some are more worthy of praise than others; fails to encourage some students
No	Modelling resilience by expressing vulnerability	Showing that it is possible to adapt and achieve despite difficulties now or in the past	"I struggled to write clearly for years, but I kept asking for feedback and got better."	Helps students to perceive the teacher as a model for coping with challenges; makes classroom more accepting space for failure
Deleted (overlapping)	Offer rewards	Offering—but not yet providing—extrinsic rewards: privileges or items that are not inherent to the task, but are provided in an effort to promote a behaviour.	"If you all finish the questions, I'll play a short video clip."	To direct behaviour so students know what behaviour the teacher wants to see

Deleted (overlapping)	Provide frequent constructive criticism	Frequently provide constructive criticism (informative feedback regarding areas of improvement)	Consistently observing the class to provide feedback for getting unstruck	Promotes continual improvement in abilities.
Deleted (overlapping)	Praise a student's fixed qualities	Provides praise that targets the talents or fixed qualities of students	"You are very smart"	Affirms students natural abilities
Deleted (overlapping)	Praise a student in public	Praise a student in public	Praise in front of the class	Generates pride within students receiving praise
Deleted (overlapping)	Provide regular praise	Regularly praise students	Teacher consistently observes the class, praising students for correct answers	Provides continual affirmation of progress and improvement
Deleted (overlapping)	Communicate in a perspective-taking way	Show that you have taken a students perspective	"I can see why you'd find this activity tricky because it's the first time you have tried it."	Communicates that teacher understands the students frame of reference
Deleted (overlapping)	Provide rewards fairly	Provide rewards when the expected behaviour is observed	"You all did your homework, so as I promised, we can watch a YouTube video today"	Adds external, tangible signal of which behaviours are desirable/valued by the teacher
No	Humour	Use authentic humour so the class is fun	"What did the triangle say to the circle? You are pointless"	Alleviates anxiety and reduces ego-involved goal-focus; increases warmth for teacher; stimulates interest
Deleted (overlapping)	Create heterogeneous groups	For group activities, assign students so that each group has a mix of abilities	"Take a playing card, and find the other students with the same suit as you"	Removes public signalling of incompetence; increases chance of balanced frames of reference and more diverse interpersonal connections

No	Creating meaning through using parables, stories, analogies, or metaphors	Use parables, stories, analogies, or metaphors to help students to connect abstract constructs into concrete examples	"mistakes are stepping stones, not stumbling blocks"; "treating others well is like sowing good seeds, eventually you'll reap a good harvest"	Promote understanding through narratives that students can connect to, and provide examples that students can follow.
Deleted (overlapping)	Set goals on behalf of the class	Teachers setting goals for students rather than providing individual challenges	"Complete all questions on page 4 by the start of class tomorrow. I want you all to get at least 9 out of 10."	Provides the same level of challenge for the whole class.
No	Provide incorrect modelling examples	The teacher unintentionally models behaviour in a way that is unlikely to produce the desired outcome.	"Helium is the first element of the periodic table"	Undermines progress because students are likely to reproduce the undesirable approach.
No	Regular communication with parents	Teachers engaging in regular contact (e.g., phone, email, text) with parents about the activities in the class,	The teacher calls a parent when she notices that a student has been particularly disengaged and unenthusiastic to talk and when the children improved, was particularly engaged and/or creative.	If contact is not transactional, it may support connections between the students' home and school life, identifying ways that key people in each domain (e.g., parents, teachers) can support each-other, building trusting relationships