Why do some schools struggle to retain staff? Development and validation of the Teachers' Working Environment Scale (TWES)

Sam Sims
UCL Institute of Education

Many school systems face recurring shortages of qualified teachers. Research suggests that improving teachers' working environment is one way of improving teachers' job satisfaction and retention in the profession. However, working environment is inconsistently measured in the literature, hampering cumulation of knowledge. This article develops the first validated teacher-report measure of working environment. Theory and empirics were synthesised to define relevant constructs, then focus groups, expert consultation and cognitive interviews with teachers were used to generate and refine items. The resulting questionnaire displayed the hypothesised factor structure in two samples of teachers. The scale also showed strong predictive validity and the results help resolve a puzzle around the importance of workload in the existing empirical literature. Through improved measurement, the scale can help identify steps by which school leaders can improve retention.

1. Introduction

There are longstanding shortages of appropriately qualified teachers in many countries (Dolton, 2020; White & Smith, 2005). Indeed, more than one third of headteachers in Luxembourg, Germany, Ireland, Estonia, Belgium, Spain, the Netherlands, Czech Republic and Hungary cite a shortage of appropriately qualified staff as hindering their capacity to provide instruction (OECD, 2018). These shortages reflect large flows of teachers out of the profession (Sims, 2018), which is particularly concerning, because high levels of turnover in specific schools is associated with reduced attainment, both through the loss of school-specific knowledge and through the wider disruption it causes (Gibbons, Scrutinio, & Telhaj, 2018; Ronfelft, Loeb, & Wyckoff, 2013). Improving teacher retention is therefore an important goal for policymakers and school leaders. Indeed, concerns about ensuring a sufficient supply of teachers in Europe are sufficient to have warranted a two-volume report on the subject by the European Commission (EC, 2013; 2014).

1.1 Conceptualising teachers' working environment

An empirical literature stretching back twenty years suggests that the quality of the working environment in a teacher's school is an important determinant of retention (Boyd et al., 2011; Ladd, 2011; Kraft, Marinell, & Yee, 2016; Sims, 2020; Weiss, 1999). In particular, the nature of school leadership experienced by teachers appears to play a particularly important role in teachers' decisions about whether to quit (Boyd et al., 2011; Ladd, 2011; Sims, 2020). Standards of pupil behaviour and the extent of staff collaboration have also been found to correlate with retention (Kraft et al., 2016; Sims, 2020). A closely related strand of the literature suggests that the same broad set of working environment variables have an important influence on teacher job satisfaction, suggesting that this might mediate the relationship with retention (Collie et al., 2016; Johnson, Kraft, & Papay, 2012; Skaalvik & Skaalvik, 2011; Reeves, Pun, & Chung, 2017; Sims, 2020). Improving teachers' working environment is therefore an important route to improving retention.

While much research has addressed teachers' working environment, the language used to label the construct differs widely, from *school working conditions* (e.g. Ladd, 2011), to *organisational/school context* (e.g. Simon & Johnson, 2015), to *professional/working environment* (e.g. Grissom, Nicholson-Crotty, & Harrington, 2014). Regardless of the label employed, the concept encompasses school policies (which are determined primarily school by leaders) and relationships (which are influenced by both leaders and teachers through their interactions) as they affect teaching staff (Papay & Kraft, 2017). The term working environment is adopted here as it does a better job than *school context* in drawing attention to the work-related aspects of teachers' lives and also avoids the ambiguity of the word *conditions* which is often used to refer to the non-pecuniary aspects of employment contracts.

More precisely, the term working environment (WE) is used here to refer to: policies and shared ways of working that are within the proximal control of school leadership and affect teachers' ability to fulfil their job roles. The term 'ways of working' is intended to capture the ways in which the organisation operates that are less formal than policies. The 'shared' prefix emphasises that these must operate across multiple members of staff to be considered part of the environment, while also allowing the concept to incorporate bottom-up aspects which may not be universal within the school. This also serves to exclude individual characteristics, such as teachers' personal sense of vocation (Gu & Day, 2013). 'Within the proximal control of school leaders' restricts the scope of the term to things that leaders can reasonably be expected to influence, thus excluding government policy or socio-economic characteristics of the school's pupil intake, while also allowing a role for teachers to have an influence via their school leadership. Finally, the condition that these must 'affect teachers' ability to fulfil their job roles' restricts the domain to things directly relevant to teachers' work educating pupils.

1.2 The present research

Unlike other correlates of teacher retention, such as pay (Hendricks, 2014), which is often constrained by government funding or regulation, teachers' working environment is within the influence of school leaders (Donaldson, 2013; Orphanos & Orr, 2014). This makes it a promising way for schools to improve retention. If research is to inform leaders' actions, however, it needs to move beyond identifying abstract correlates of improved retention such as 'leadership' to identify "managerially controllable variables" which have "direct action implications" (Shrivastava & Mitroff, 1984; p.23). This article aims to contribute to this process by developing the first validated questionnaire measuring concrete aspects of working environment in schools relevant to teachers' job satisfaction and retention. This should help bring enhanced conceptual clarity to the literature and ensure that findings cumulate to provide a steadily improving body of knowledge on which school leaders can base their decisions regarding how to improve the working environment in their schools. In sum, the paper develops a Teacher Working Environment Scale (TWES), which aims to improve both the validity and the utility of the way in which teachers' working environment is measured.

2. Review of measurement

This section reviews the approach taken to measurement in the existing literature. Sub-section 2.1 reviews the instruments used in prior research, sub-section 2.2 then summarises two important limitations of these approaches, finally sub-section 2.3 considers whether there are existing instruments elsewhere in the literature that could be adapted, instead of developing a new instrument from scratch.

3

2.1 Measurement in the existing literature

Instruments used in the existing literature can be grouped into three categories. The first of these comprises questionnaires developed by government agencies. Early research in this tradition tended to use the teacher questionnaire from the US Federal Government's *Schools and Staffing Survey* (SASS; Ingersoll, 2001; Shen, 1997; Weiss, 1999). The contents of the questionnaire changed across survey waves[1] but tended to include a group of questions on 'perceptions and attitudes towards teaching', which included several items capturing working environment. More recently, researchers have employed instruments developed by state or city authorities such as Chicago (Allensworth, Ponisciak, & Mazzeo, 2009), North Carolina (Ladd, 2011), Massachusetts (Johnson, Kraft & Papay, 2012), and New York City (Kraft, Marinell, & Yee, 2016; Marinell & Coca, 2013). While the exact content of these instruments varies, there is some similarity as a result of the North Carolina Working Conditions Survey being used as the basis of other state-level surveys, such as the Massachusetts Teaching and Leading survey (MassTeLLS). However, there is very little published evidence on the validity and reliability of any of these survey instruments.

A second set of instruments used in the existing literature are those developed for a specific research project. For example, Loeb, Darling-Hammond and Luczak (2005) analysed data from an apparently proprietary phone survey in California. Similarly, Boyd et al. (2011) compiled their own questionnaire – again drawing some questions from the *Schools and Staffing Survey* – for use in their survey research in New York. Skaalvik and Skaalvik (2011) also employ a range of novel scales in their study of working environment and teacher job satisfaction in Norway. In all three cases, very little evidence was presented for the validity of the instruments used. This makes it hard to assess the relationship between the data derived from these instruments and the constructs of interest to researchers working in this literature.

The third set of instruments used in the existing literature are the questionnaires from the 2008, 2013 or 2018 waves of the Teaching and Learning International Survey (TALIS; Duyar, Gumus, & Sukru Bellibas, 2013; Fackler & Malmberg, 2016; Sims, 2020). TALIS is an international survey that employs a centrally determined questionnaire, which is subsequently translated into the language of participating countries. Like the SASS, the exact contents of the TALIS questionnaire vary from wave to wave[2], though the instrument tends to contain around 40 items related to working conditions (Sims, 2020). The original TALIS questionnaire was developed for the purposes of the survey using expert consultation (OECD, 2010) and psychometric evidence for the reliability and validity of the various sub-scales has accumulated with each survey sweep

---

[1] Instruments for all survey sweeps can be accessed here https://nces.ed.gov/surveys/sass/
[2] http://www.oecd.org/education/school/talis2018questionnaires.htm

4

(OECD, 2014a). However, a critical shortcoming of the TALIS questionnaire for the purposes of measuring teachers' working environment is the lack of a validated scale for measuring supportive leadership, which existing research suggests is the most important aspect of the working environment construct (Boyd et al., 2011; Johnson et al., 2012; Kraft et al., 2016; Ladd, 2011; Sims, 2020; Weiss, 1999).

2.2 Limitations of measurement in the existing literature

This brief review highlights two limitations in the existing literature. The first and most fundamental problem with existing measurement practices is that they leave the reader unable to assess construct validity (Cronbach & Meehl, 1955; Flake & Fried, 2019). That is, without carefully defining the construct, we cannot assess whether a set of survey items capture it appropriately (Shadish, Cook, & Campbell, 2002). This is a particular problem in exploratory factor analysis, where the reader is very often given a label for a factor e.g. 'leadership' but no definition of this construct. Whether or not a set of items fully or appropriately represent a label is indeterminate; whereas the extent to which a set of items represent a well-defined construct can be critically evaluated. In addition, school leaders reading the research are likely left wondering what exactly a label such as 'leadership' really means. The practical implications of the findings thus remain opaque. This problem affects much of the existing literature (Ingersoll, 2001; Kraft, Marinell, & Yee, 2016; Ladd, 2011; Kraft, Marinell, & Yee, 2016; Sims, 2020; Weiss, 1999). Addressing these limitations is therefore an important next step in the development of this literature.

The second problem is that a very wide variety of methods have been used to measure teacher working environment. Indeed, almost every paper uses a unique instrument. This makes it difficult to compare findings across papers and limits the potential for cumulative knowledge building as the literature develops. It might be assumed, for example, that the factor labelled 'leadership' in each of the above studies is the same factor. Yet if each of these 'leadership' factors is in fact measured by a different (and perhaps entirely non-overlapping) set of survey items, then they may in reality be measuring quite different constructs – a misapprehension referred to in the psychology literature as the *jingle fallacy* (Thorndike, 1904). Conversely, researchers might be using similar sets of items from a popular survey instrument, such as the MassTeLLS, but giving the factor a different label. Mistaking these for two different constructs has in turn been dubbed the *jangle fallacy* (Kelley, 1927). Developing the first validated questionnaire makes it more likely that, in future, researchers will cohere around the use of this valid instrument. This will help new research findings cumulate into a coherent body of knowledge.

2.3 Could existing scales be adapted, rather than starting from scratch?

Are there existing instruments that could be adapted, rather than developing a new instrument from scratch? One plausible candidate is the Areas of Worklife Scale (AWS; Leiter & Maslach, 2003), which uses 16 items to measure six main predictors of burnout (Leiter & Maslach, 1999). Two of the AWS factors – workload and community – have also been shown to predict job satisfaction and retention among teachers (Kraft et al., 2016; Skaalvik & Skaalvik, 2011) and the AWS has previously been administered to teachers (Timms, Graham, & Cottrell, 2007). Despite this, the AWS has an important limitation for measuring teachers' working environment in that it does not include measures of school leadership and discipline, both of which have been found to be important for teacher retention (Kraft et al., 2016; Sims, 2020).

The closest alternatives within the education literature are scales designed to measure school climate. Careful reviews of such instruments have identified several well-validated questionnaires for measuring this construct (Clifford et al., 2012; Gangi, 2009). The difficulty with school climate, however, is that it is far broader in scope: "School climate is based on patterns of people's experiences of school life and reflects norms, goals, values, interpersonal relationships, teaching and learning practices, and organizational structures" (National School Climate Council, 2007). School climate also tends to be conceptualised as being shared by, and relevant to, all members of a school community including teachers, support staff, parents and pupils (Thapa et al., 2013). Teachers' working environment, by contrast, encompasses perhaps only two of the six aspects of school climate (interpersonal relationships and organisational structures) and applies solely to teachers.

In summary, while there are several well-validated scales that measure neighbouring constructs in the organisational and educational psychology literature, they all miss out on important aspects of the working environment construct. Moreover, the constructs they do measure might not be suitably focused on teachers, as opposed to other members of the school community. Taken together with the evidence from previous sections, this suggests that there is a case for developing and validating a new teacher working environment scale (TWES). The rest of the paper reports on a programme of research that sought to achieve this goal.

3. Methodology

As with all scale development, my aim is to create a valid measure, such that variations in teachers' working environment causally produce variations in the TWES score (Borsboom, Mellenbergh, & van Heerden, 2004). This will be achieved through a process of construct validation, which involves "integrating evidence to support the meaning of a number which is assumed to represent a… construct of interest" (Flake et al., 2017). As psychometrics has

progressed as a field, different 'types' of, or tests for, validity have developed – including constructs, content, predictive, convergent and divergent validity (Cronbach & Meehl, 1955; Campbell & Fiske, 1959; Strauss & Smith, 2009). However, as Loevinger (1957) argued, construct validity is the essence of the idea of validity. This section provides a general overview of the methodological approach, which is conceptualised as falling in three distinct phases: substantive (section 4), structural (section 5) and external (section 6) (Benson, 1998; Loevinger, 1957). As multiple research methods are used in each phase, details of the participants, procedure and analytical approach are provided in each of the subsequent sections.

The substantive phase involves specifying the theoretical domain of, and providing an operational definition for, the construct. This involves careful review of theory and empirics from the literature in order to specify how measurable aspects of the world can be used to capture the construct of interest. The ultimate motivation for developing the TWES is to help provide actionable insights for school leaders looking to improve teacher job satisfaction and retention. The TWES therefore aims to measure concrete aspects of teachers' working environment, such as teachers offering each other advice, as opposed to abstract psychological principles such as 'connectedness'. Psychological theory will instead be used as a foundation to inform scale development. The word foundation is particularly apt here because while foundations are integral to the structure of a building and their presence and outline is reflected in the external walls of the building visible above ground, they cannot be directly observed. In the same way, the items chosen will be structured around and reflective of the underlying theory, but the theoretical constructs will not be directly visible in the scale.

The structural phase involves using psychometric methods to tests the fidelity with which the observed relationships between the items and factors reflect those hypothesised in the prior, substantive phase. This requires collecting data using a draft version of the instrument to see whether the data are indeed consistent with the theory. Do the set of items that are intended to represent one construct correlate closely with each other? Do the hypothesised set of constructs emerge from the data? It should be noted that, while useful, all of these tests are indirect, in the sense that they do not directly establish whether changes in teachers' working environment causally produce variation in TWES scores (Borsboom et al., 2004). Psychometric tests without the preceding substantive work are thus insufficient for demonstrating validity (Maul, 2017). More generally, this article followed the advice of Gehlbach & Brinkworth (2011) to 'measure twice and cut once'. That is, the emphasis is placed on carefully conceptualising and operationalising working environment in the substantive phase; as opposed to emphasising cutting out items with undesirable psychometric properties in the structural phase, which runs the risk of compromising validity by undermeasuring certain aspects of the construct.

The third, external phase involves establishing the relationship between the TWES score and the constructs that it is supposed to predict. Does a higher score correlate positively with higher job satisfaction and lower turnover? This stage provides another useful check. It is, however, even more indirect as a test of the validity of the scale in that it provides only circumstantial evidence on the causal relationship between the construct and the score. It is worth noting that, in effect, most or all of the existing research on teachers' working environment has started at the external phase of the process, skipping the first two.

A final general point about the approach is that 'Strongly agree' to 'Strongly disagree' response scales are not used in a blanket way across all items. This reflects empirical evidence that using only agree/disagree response scales can have undesirable measurement properties (DeCastellarnau, 2018). In particular, they should be avoided when the item aims to establish a degree or an extent (Fowler, 2009; Krosnick, 1999). For example, if a question aims to establish how clear the processes for dealing with behaviour in a school are, then the response scale should specify different levels of clarity, rather than looking for levels of agreement/disagreement with a statement that they are 'clear'. Avoiding the blanket use of agree/disagree scales requires a more thoughtful approach, tailored to the specifics of each item (DeCastellarnau, 2018).

4. Substantive Phase

4.1 Theorising teachers' working environment

Valid measurement cannot occur without first theorising the constructs of interest (Borsboom et al., 2004). The two psychological theories that have been tested most extensively in relation to teacher turnover and associated psychological states are the Job Demands-Resources model (JDR) and Self Determination Theory (SDT). For the purposes of teaching, JDR states that if job demands (aspects of work that require sustained psychological or cognitive effort) exceed job resources (aspects of work which are functional for achieving work goals) then teachers will become disengaged from and ultimately leave their job (Demerouti et al., 2001). Importantly, recent empirically driven revisions of the framework have clarified that it is primarily when job demands are perceived as hindrances (e.g. conflict, ambiguity, politics and red tape) that they have this negative consequence (Crawford, Lepine, & Rich, 2010). The JDR model has been shown to fit data on teachers in Germany (Feuerhahn, Bellingrath, & Kudielka, 2013), Finland (Hakanen, Bakker, & Schaufeli, 2006) and Spain (Prieto, Soria, Martinez, & Schaufeli, 2008).

SDT states that humans need access to three basic psychological needs in order to be energised and motivated: autonomy, competence and connectedness (Deci & Ryan, 2000; Deci & Ryan, 2008). The theory predicts that where these basic psychological needs are not met at work, individuals will seek to change their job in order to access them. It should be noted that autonomy

here means doing something because you endorse the reasons behind it, rather than simply being left to your own devices. SDT has also been shown to fit well with data on workers in general (Gagne & Deci, 2005) and teachers in particular (Collie, Shapka, Perry, & Martin, 2016; Reeve & Su, 2014).

Although the two theories were developed separately, they are mutually compatible. The JDR model can be thought of as describing a lower bound on what individuals can put up with at work and the SDT framework as specifying what helps individuals thrive at work, and what they strive for in choosing employment. The two theories can also be seen as complementary, with aspects of working environment that constrain autonomy or make it harder for teachers to do their job in a highly competent manner falling within the category of demands; while aspects of working environment that help teachers perform competently or that provide social support falling within the category of resources. Indeed, recent empirical research has integrated the two frameworks in research on teachers' engagement at work, providing evidence for their complementarity (Fernet et al., 2015; Fernet et al., 2013). Both theories will be reflected in the scale development described below.

4.2 Theorising the subscales

Careful qualitative research suggests that there are several distinct aspects of teachers' working environment (Johnson & Birkeland, 2003) and this has been borne out in subsequent research using factor analysis to explore quantitative data (Boyd et al., 2012; Kraft et al., 2016; Ladd, 2011; Sims, 2020). This implies that a scale measuring teachers' working environment will consist of a set of sub-scales, each measuring a different aspect of working environment. Which sub-scales should be included in the TWES? The remainder of this section assesses a number of options based upon the theory introduced above and existing qualitative and quantitative empirical research.

Qualitative work consistently highlights the importance of school leadership for teacher retention (Johnson & Birkeland, 2003; Johnson et al., 2004; Kraft et al., 2015; Perryman & Calvert, 2019). This has been strongly corroborated by quantitative research in the UK, Norway and elsewhere, which has consistently found leadership to be an important predictor of teacher retention (Boyd et al., 2011; Johnson et al., 2012; Kraft et al., 2016; Ladd, 2011; Sims, 2020; Weiss, 1999) and teacher job satisfaction (Johnson et al., 2012; Skaalvik & Skaalvik, 2011; Sims, 2020; Weiss, 1999). The importance of leadership is also readily interpretable within the JDR and SDT frameworks in that leaders can either create autonomy-constraining demands on teachers or provide resources in the form of e.g. advice and mentoring (Salokangas et al., 2020; Steen-Olson & Eikseth, 2010). The specific language used to describe this factor differs across articles, countries and time periods

(Simkins, 2012). Herein, this construct will be referred to as *Supportive Leadership* on the grounds that this is aligned with the JDR theory, captures the essence of the range of constructs addressed in the existing literature, and (unlike 'administration') has a common meaning across English-speaking countries. Supportive Leadership is defined as the exercise of influence and direction setting in order to help teachers achieve their work goals (Bakker & Demerouti, 2017; Cuban, 1988). Established measures of specific 'types' of leadership (e.g. transformational leadership [Burns, 1978]) have purposely not been used here, because the objective is to develop a much narrower bandwidth measure focused only on the way in which leaders affect teachers' working environment.

Detailed qualitative accounts also tend to highlight pupil behaviour as being an important aspect of working environment (Johnson & Birkeland, 2003; Johnson et al., 2004; Kraft et al., 2015). This is directly relevant to teachers' working environment because poor behaviour directly inhibits teachers' *ability to fulfil their job roles*. Indeed, the importance of behaviour is straightforward to conceptualise within the JDR and SDT frameworks in that poor behaviour is a potentially serious demand on teachers, which is likely to inhibit their sense of competence. Quantitative studies from the UK, Norway, Portugal and elsewhere have also found evidence that behaviour is predictive of both teacher job satisfaction (Lopes & Oliveira, 2020; Skaalvik & Skaalvik, 2011; Weiss, 1999) and retention (Kraft et al., 2016; Sims, 2020). The literature is not entirely consensual however, in that Boyd et al. (2011) find no relationship between discipline and teacher retention. Herein, this construct is referred to as Behaviour Policy, rather than simply behaviour, because this captures the *shared* aspects within the *proximal control of leaders*. It is defined as the "policies and actions taken by school personnel to prevent or intervene with unwanted behaviours" (Cameron, 2006, p291).

Qualitative research from the UK, Greece, Sweden and elsewhere emphasises the importance of teacher relationships/collaboration for retention (Johnson & Birkeland, 2003; Johnson et al., 2004; Kraft et al., 2015; Perryman & Calvert, 2019; Toropova et al., 2021). Interestingly, the results from quantitative research are less consistent. Some studies find relationships are predictive of job satisfaction (Johnson et al., 2012) while others do not (Sims, 2020). Similarly, some find that staff relationships are associated with retention (Johnson et al., 2012; Kraft et al., 2016) while others do not (Sims, 2020; Boyd et al., 2011). Looking across the quantitative studies, it is clear that this construct (if indeed it is a single construct) has been labelled in many different ways, ranging from 'staff relations', to 'colleagues', 'collaboration', 'culture' and 'cooperation'. The items measuring the constructs also vary considerably. This raises the possibility that inconsistent findings are the result of jingle or jangle fallacies. The JDR and SDT theories suggest that relationships/collaboration may help if they provide useful resources such as advice or bolster a sense of connectedness. However, enforced collaborative activities might also constitute a hindrance demand (Hopkins et al., 2019). In the TWES, this construct is referred to as Collegiality, which captures the sense of teachers working

together and supporting each other socially and emotionally, while emphasising the quality of relationships over particular collaborative activities (Jarzabkowski, 2003; Kelchtermans, 2006).

Qualitative research in the UK, Iceland, Ireland, Norway and elsewhere has strongly emphasised the importance of workload or time shortages for teacher retention (Barmby, 2006; Cooper, Gibson Research, 2018; Jóhannesson, 2006; Johnson & Birkeland, 2003; Perryman & Calvert, 2019; Steen-Olson & Eikseth, 2010; Torres, 2016). However, the evidence from quantitative research again provides a less consistent picture. Some studies find that workload is predictive of retention and job satisfaction (Johnson et al., 2012; Skaalvik & Skaalvik, 2017) while others do not (Boyd et al., 2011; Ladd, 2011; Sims, 2020). As with collegiality, theory can potentially help resolve these conflicting findings. The JDR model, particularly the revised version, emphasises that it is not the overall quantity of work that matters. Rather it is whether the specific tasks are perceived as hindrances by teachers (Crawford et al., 2010) and thus undermine their autonomy. The terms 'workload' and 'time' seem too broad and too quantitative to capture this notion. Herein, this construct is referred to as Compliance, which is defined as the need to go *out of ones way* in order to follow a procedure imposed by a superior (Bozeman, 1993).

Other constructs that have appeared in the teachers' working environment literature but were not included in the development of the TWES are influence/autonomy, professional development and facilities. Influence and autonomy are not included as a separate factor here because they are conceptualised, and indeed often appear in exploratory factor analyses as, part of the leadership/management construct (e.g. Boyd et al., 2011; Sims, 2020). Professional development – in the formal, organised sense – makes very little appearance in the qualitative literature and, where it appears in the quantitative research, it is generally not found to be related to turnover (Boyd et al., 2011; Ladd, 2011; Sims, 2020). Facilities – in the sense of the physical infrastructure of the school – also appears in a number of papers but are not found to be correlated with either job satisfaction or retention.

4.3 Focus groups for developing the items

Participants. In order to develop items for the four subscales, two focus groups, each containing six teachers, were held. Participants included current and former teachers in order to incorporate the views of both those who have and those have not left the teaching profession. Participants were recruited through a snowballing approach using networks at a teacher training institution in the UK. All participants gave their consent to participate.

Procedure. The focus groups were not intended to identify or theorise relevant constructs; this had already been achieved in a prior phase. Rather, the focus groups aimed to capture "how participants think about the focal construct[s] in their own words, with minimal prompting"

(Gehlbach & Brinkworth, p. 382). The participating teachers were asked to reflect on schools they had worked in, with both good and bad working environments, and then left to talk. Where particular comments or turns of phrase closely reflected the theoretical framework set out above, this was used to develop an item. For example, two participants used the word 'approachable' to describe the leadership in schools with good working environments, explaining that this helped support good communication and allowed them to develop strong working relationships. I therefore developed the item "School leaders are approachable". In order to ensure that each of the four constructs was fully represented by the set of questionnaire items, wording or phrasing from the focus group discussions were collected that reflected each of the components of the underlying theoretical framework (see Appendix Table 1).

Results. Under Supportive Leadership, the following items were developed: "School leaders provide opportunities for teachers to participate in decision making" (demands/autonomy) "School leaders recognise teachers for doing a good job" (resources/competence); and "School leaders are approachable" (resources/connectedness). For Behaviour Policy, items were developed reflecting resources/competence e.g., "The behaviour policies allow me to get on with teaching" and reflecting resources/connectedness e.g. "Middle and senior leaders can be trusted to support me in relation to pupil behaviour". The focus group provided no examples relating to autonomy with behaviour policy and, on reflection, it was concluded that autonomy was indeed theoretically less relevant to the behaviour policy construct. For Collegiality, items were developed reflecting resources/competence e.g., "The teachers in my department/phase work together to solve problems" and resources/connectedness e.g. "I have colleagues who care about me." Again however, on reflection, it was concluded that autonomy was not so theoretically relevant to the Collegiality construct. Finally, for Compliance, items were developed reflecting demands/autonomy e.g., "I am expected to do tasks which do not contribute to pupils' education" and demands/competence e.g., "I am asked to teach subjects or age-groups for which I have not been trained". With respect to Compliance, connectedness seemed theoretically less relevant and no items pertaining to this component of theory were developed. Appendix Table 1 contains a summary.

4.4 Expert consultation for refining the items

Next, the draft questionnaire was sent to six domain experts, who were selected based on having previously published research on teachers' working environment. They were asked to comment on the items using a structured feedback form. The first page of the form elicited comments on the clarity of the items and appropriateness of the proposed response scale. This resulted in several changes, including some of the items shifting from an agree/disagree response scale to a frequently/infrequently response scale. The second page of the structured form asked the

consultees to assess whether any important aspects of the constructs were undermeasured by the proposed items. This resulted in further changes, such as the addition of the item "Marking gets in the way of teaching". This process resulted in 26 items across the four constructs (Appendix Table 1).

4.4 Cognitive interviewing for refining the items

Participants. In order to further refine the items, six cognitive interviews (Willis, 2004) were conducted, each with a single teacher. Again, participants included a balance of current and former teachers and were recruited through networks at a teacher training institution in the UK. Participants all gave their consent to participate and for the conversations to be recorded.

Procedure. The aim of this process was to understand whether respondents understood and interpreted the items in the way they were intended. To do this, the interviewees were asked to respond to the draft questionnaire orally, talking through their thought process as they did so. Probes were then used to check how respondents had interpreted important words. Where ambiguities or misunderstandings were identified, the item was amended before the next cognitive interview, resulting in an iterative process of refinement.

Results. This process resulted in a number of improvements to the items. For example, when responding to the item "I understand why decisions have been made" (Supportive Leadership) one interviewee noted that they often thought they knew why, but this was largely due to their own inferences. As such, the item was revised to read "Leaders clearly communicate the reasons for which decisions have been made." Another example of a revision relates to the item "I am asked to do things to generate evidence of school improvement." One participant questioned the use of the word 'ask' on the grounds that, although they had never been explicitly asked to do so, they knew they were expected to. A subsequent participant also pointed out that teaching passively generates evidence of school performance through e.g. the normal course of examination. The final revised version of the item therefore read "I am expected to do things solely for the purpose of generating evidence". A third example relates to the item "School leaders say thank you to teachers for the work they do". After a participant pointed out that they could not really vouch for whether leaders said thank you to *other* teachers the item was revised to "Leaders say thank you to me for the work that I do". Several other changes, including further refinements to response scales, were made as a result of this process. The full set of changes can be seen by comparing Appendix Table 1 with the final instrument in Appendix Table 2.

5. Structural Phase

5.1 Exploratory factor analysis

Participants. In order to test the structure of the questionnaire, the revised version of the instrument was administered in three secondary schools in England in the 2018/19 academic year. Headteachers of these schools were invited to take part via an email distributed to members of a professional association in the UK. With the headteachers' consent, schools provided email addresses for all teaching staff. This yielded 236 valid addresses, all of which were contacted and invited to take part in an online survey using Smart Survey. Of these, 169 (72%) gave their consent to participate and provided an anonymous survey response (69% female; modal age category 30-34 years old).

Analysis. The resulting data was coded so that a higher score always represented 'better' working environment. Less than 1% of the data was missing for any given variable, rendering imputation unnecessary (Cheema, 2014). The polychoric correlation matrix between the 26 survey items was then constructed and subjected to exploratory factor analysis (EFA) using an oblique (promax) rotation. Following Flake et al. (2015), the decision about how many factors to extract was based on three sources of evidence: eigenvalues > 1, the minimum average partial procedure and parallel analysis. Internal consistency was explored using alpha coefficients and item-rest correlations.

Results. Eigenvalues and the minimum average partial procedure both suggested four factors be extracted. Parallel analysis was technically indeterminate in that the eigenvalues from the factor analysis never dropped below those from the parallel analysis (Figure 1). Having said that, the two lines become extremely close – and then remain so – after the fourth factor, indicating that the eigenvalues from the fifth factor onwards are not materially higher than the fifth factor extracted from a correlation matrix of simulated random values. Hence, in line with the factor structure originally hypothesised, the first four factors were retained.

<Figure 1>

Table 1 shows the resulting factor pattern matrix. All items load > |0.3| on the factors on which they were hypothesised to, providing evidence consistent with the construct validity of the scale. The pattern of loadings is also very simple: only three of the 26 variables cross-load on other factors and, of these, all three load more strongly on their hypothesised factor. Table 2 shows the item wording, coefficient alpha and item-rest correlations for each factor or item. The alphas (α) are high – ranging between 0.81 and 0.94, which indicates strong internal consistency for each subscale. The item-rest correlations (IRC) are also generally high: ranging from 0.52 to 0.84. The one exception to this is item 4 in Compliance, "I am asked to teach subjects or age-groups for which

I have not been trained" which has an IRC of 0.34. Although this is low, the item was not cut from the scale on the grounds that it is consistent with JDR theory (teaching out-of-subject is likely to be experienced as a substantial demand) and SDT theory (it is likely to impinge on teachers sense of competence). There is also direct empirical support for the relationship between teaching out-of-subject and retention, at least for early-career teachers (Donaldon & Johnson, 2010). Removing it would therefore likely undermine content validity of the TWES (Flora & Flake, 2017).

<Table 1>

<Table 2>

5.2 Confirmatory Factor Analysis

Participants. In order to provide a further, independent test of the factor structure, a second dataset was collected from a separate convenience sample of 24 schools in England in the 2019/20 academic year, using the same questionnaire. These schools were again recruited by email to headteachers through a professional association. The participating schools provided email addresses for all teaching staff, yielding 1,839 valid addresses, of which 1,233 (67.4%) consented to participate and provided an anonymous response (64% female; modal age category 30-34 years old). Missing data was again below 1% for all variables.

Analytical approach. A set of confirmatory factor analyses (CFA) were conducted in order to further test the hypothesised factor structure of the instrument. Specifications including different numbers of factors (subscales), as well as more complex factor structures. These tests were conducted using a robust weighted least squares (WLSMV) estimator to accommodate the categorical nature of the data. For each specification, the model-data fit was assessed using a range of fit statistics: the comparative fit index, with acceptable fit indicated by CFI>0.095; the Tucker–Lewis index, with acceptable fit indicated by TLI>0.95; the root mean square error of approximation; with acceptable fit indicated by RMSEA<0.06; and the standardized root mean square residual, with acceptable fit indicated by SRMR<0.08 (Hu & Bentler, 1999).

Results. The first row of Table 3 shows the results from a simple one-factor model in which all of the observed variables load on an overall teacher working environment factor. The model fit is weak, which again suggests that the TWES might contain multiple subscales measuring distinct factors. Since Figure 1 suggested that three, four and five-factor models were the most plausible, the next three rows of Table 3 reports test of fit for these three specifications, with each factor measured using the manifest variables suggested by a promax rotation in each case. The three-factor model (Behaviour Policy, Collegiality, and a combined Leadership/Compliance factor) shows much better fit than the single-factor model, though only one of the three fit statistics meets conventional criteria for acceptable fit. The four-factor model shows notably better fit than the three-factor model, with

two of the four fit statistics meeting the conventional criteria for acceptable fit, and the TLI being right on the cut-off (0.95). The five-factor model (which splits the Collegiality factor into two) shows similar fit to the four-factor model. As the four-factor model is more parsimonious, was suggested by the exploratory factor analysis and is better aligned with theory (Flora & Flake, 2017), the four-factor model was concluded to be preferable to the five-factor model.

<center>**<Table 3>**</center>

Recall that each of these four factors is theorised to be part of a set of constructs related to teachers' working environment and, consistent with this, Table 4 shows that the four factors are correlated with each other. This raises the question of whether there is there a separate overall factor for teachers' working environment and, if so, how do the first four factors relate to it (Gustafsson & Aberg-Bengtsson, 2010)? The last two rows of Table 3 hence report on more complex models investigating the nature of this relationship. Row five reports fit statistics from a higher-order factor model (see Figure 1C in Brunner, Nagy, & Wilhelm, 2011; or Appendix Figure 1) that includes a higher-order latent variable for Teachers' Working Environment that is itself measured by the four lower-order latent variables: Supportive Leadership, Behaviour Policy, Collegiality and Compliance. The model does not show clearly better fit to the data than the four-factor model. This suggests that, while some of the pairwise correlations between the factors are fairly high (Table 4), there is little variance that is shared across all four. The fifth row of Table 3 reports fit statistics on a nested factor model (See Figure 1D in Brunner, Nagy, & Wilhelm, 2011; or Appendix Figure 2) that again augments the model with an additional latent variable, but this time the latent variable is measured by all the observed questionnaire variables, rather than the four latent variables. This is the best fitting model, with only the RMSEA (0.061) falling marginally short of the conventional cut-off (0.06). This implies that the questionnaire items do all share common variance but that the four factors account for variance above and beyond that shared by the observed variables in general (Chen et al., 2012). Taken together, this evidence is consistent with the idea that there is a broad-scope (general) latent variable representing Teachers' Working Environment within which are four narrow-scope (specific) latent variables addressing distinctive yet pairwise-correlated aspects of working environment.

<center>**<Table 4>**</center>

6. External Phase

If the TWES scores display the hypothesised relationship with the working environment construct then we would expect the TWES score to correlate with other constructs (e.g. turnover

<center>16</center>

intentions) predicted by theory to be affected by working environment. This section provides a test of this hypothesis using the same sample employed in section 5.

Analytical approach. Table 5 shows results from regressions of the four TWES subscales on job satisfaction (Columns 1-2), intentions to leave the school (Columns 3-4) and intentions to leave (attrite from) the teaching profession altogether (Columns 5-6). Job satisfaction is measured with a single item "All in all, I am satisfied with my job", turnover intentions are measured using a single item "During the last term, I have seriously considered leaving this school" and attrition intentions are measured by a single item "During this last term, I have seriously considered leaving teaching altogether". All three items employ a four-point 'Strongly disagree' to 'Strongly agree' response scale. The models are hence estimated using ordered logistic regression and the coefficients reported in the table are odds ratios. Columns 1, 3 and 5 show the association between each TWES subscale score and the outcome, when each TWES subscale is entered into the model separately. Columns 2, 4 and 6 show the associations when all four TWES subscale scores are entered simultaneously, along with a school fixed effect term. Standard errors have been clustered at the school level to reflect the nested structure of the data. Table 6 uses reports of the working environment reported by focal teachers' colleagues to predict the job satisfaction, turnover and attrition intentions of the focal teacher. If - as hypothesised - the TWES genuinely captures variation in the shared working environment in a school, then the regression coefficients should have a similar sign and magnitude in these models. The models in Table 6 also help guard against common source bias (Podsakoff, MacKenzie, & Podsakoff, 2012).

Results. When the TWES subscales are entered separately in columns 1, 3 and 5, all four of the working environment factors show a strong association in the expected direction with all three outcomes. When the TWES subscales are entered together in columns 2, 4 and 6, a similar pattern of associations holds. With respect to job satisfaction, the odds ratios in column 2 range between 1.5 and 2.76, which implies that a one standard deviation increase in each of the factor scores is associated with a 50%-176% increase in the odds of being one category higher in terms of reported job satisfaction. For turnover and attrition the questions are phrased negatively, so that an odds ratio below one indicates a relationship in the hypothesised direction. Coefficients in Model 5 range between 0.72 and 0.56, which implies that a one standard deviation increase in the working environment factors is associated with a 28%-44% reduction in the odds of being one category higher on the turnover intentions question. For attrition, the odds ratios range between 0.8 and 0.6. As would be expected, the associations are slightly weaker for attrition intentions than for turnover intentions.

<center>**<Table 5>**</center>

In Table 6, the focal teacher's colleagues' reports of the working environment factors are used to predict the focal teacher's outcomes (satisfaction, turnover intentions, attrition). Odd numbered columns use reports of working environment from all other respondents in the focal teachers' school. Even-numbered columns use reports of working environment from all respondents in the same department (e.g. maths) as the focal teacher. The latter models have smaller sample sizes because not all respondents have colleagues in their department that also respond. Here, the results diverge depending on the specific TWES subscales. For Supportive Leadership and Behaviour Policy, the coefficients remain stable in direction and broadly comparable in magnitude. This provides reassurance that the results are not being driven by common source bias and that the TWES is capturing genuine shared aspects of the working environment. For collegiality, the coefficients have different signs in the odd and even columns of Table 5. This likely reflects the fact that collegiality shows a stronger intra-cluster correlation at department-level than at school-level (see Table 4), which it also theoretically intuitive. Indeed, in the even-numbered columns, the coefficients are of the same sign and broadly comparable magnitude with those from in Table 5. For compliance, the coefficients have the same sign and magnitude as those in Table 5 in the models predicting job satisfaction and attrition intentions, but not turnover intentions.

<Table 6>

7. Discussion

The study of teachers' working environment has made important theoretical and empirical advances in recent years, to the extent that measurement has now become an important limiting factor in new discoveries. This issue cannot be overlooked since, without evidence for the validity of the measurements employed, the numbers collected by researchers bear an unknown resemblance to the constructs of interest (Borsboom et al., 2003). In addition, findings will tend to remain disorganised and may fail to cumulate into a coherent body of knowledge (Weidmeann et al., 2017). I set out to relieve this constraint on new discoveries by developing and validating a measure of teachers' working environment. Drawing on theory and prior empirical results, consultation with teachers and domain experts, and the results of psychometric tests in two independent samples, I developed the Teachers' Working Environment Scale (TWES): a quantitative measure of policies and shared ways of working which are within the proximal control of school leadership and affect teachers' ability to fulfil their job roles.

The psychometric tests reported in the structural phase of this research provided strong indirect empirical support for the validity of the TWES by showing that the scale contained the theorised latent variables and that the questionnaire items loaded on the latent variables as

hypothesised. The regression analysis reported in the external phase provided further indirect empirical support for the validity of the TWES by showing that all four of the factor scores predicted job satisfaction, turnover and attrition intentions in line with the original hypotheses. In addition, while the intra-cluster correlations of the four working environment variables are lower than expected, the results based on colleague-report in Table 6 provide reassurance that the TWES is capturing genuinely shared aspects of teachers' working environment – as hypothesised. These tests are, however, indirect as the validity of the instrument depends fundamentally on whether changes in teachers' working environment causally produce changes in TWES scores (Borsboom et al., 2004; Maul, 2017). The substantive phase therefore provides the primary warrant for the validity of the TWES. Indeed, in line with the 'measure twice and cut once' principle (Gehlbach & Brinkworth, 2011) careful attention to synthesising existing theory and empirical findings in the substantive phase of this research appears to have minimised the need for making data-driven modifications of the scale in the structural phase.

Evidence presented on the predictive validity of the TWES also illustrates the value of improved measurement and thus the main contribution of this paper. As previously discussed, the existing literature contains conflicting findings on whether workload is related to retention. Qualitative research strongly emphasises the importance of workload for explaining declining retention (Barmby, 2006; Cooper Gibson Research, 2018; Perryman & Calvert, 2019). However, survey evidence suggests that overall hours worked by teachers in these countries has been stable in recent years (Allen et al., 2020). Similarly, while some quantitative research finds workload is related to retention (Johnson et al., 2012; Skaalvik & Skaalvik, 2017) other papers find no relationship (Boyd et al., 2011; Ladd, 2011; Sims, 2020). These conflicting findings present something of a puzzle. However, by carefully conceptualising workload drawing on the underpinning psychological theory, the TWES incorporates a new measure of 'Compliance'. This latent working environment variable was shown to be predictive of job satisfaction, turnover intentions and attrition intentions (Table 5). While further research is needed to fully resolve the issue, these substantive findings show how improved measurement via the TWES can help to resolve puzzles in the existing literature.

7.1 Limitations

The findings of this research should, of course, be interpreted with regard to its limitations. Three in particular stand out.

First, the item development and psychometric work has taken place in the English context. The extent to which the TWES is valid in schools outside England therefore remains untested. Ultimately, the applicability of the scale in other countries and further afield is an empirical question, which requires that the TWES be deployed and tested in other settings. International

comparability of findings will require formal testing of measurement invariance (Meredith & Teresi, 2006). A related point is that both of the surveys reported here were collected from convenience samples, comprised of secondary school teachers. The psychometric properties of the TWES are therefore yet to be established in primary settings. Having said this, the scale is rooted in psychological theory about the underlying nature of human motivation and persistence at work, and in that sense it should be broadly applicable (Deci et al., 2001; Ryan & Deci, 2011; Slemp et al., 2018).

A second important limitation is that there may be other subscales that should be included in the TWES. Professional development is one candidate. Kraft et al. (2016), for example, find an association between a factor they label "Leadership and professional development" and retention. However, it is not clear whether it is the leadership, or the professional development, or both, which are driving the association (see Robinson et al., 2008 for a discussion of the relationship between leadership and professional development in schools). On the basis of the full range of existing evidence, it was deemed that there was insufficient empirical support for including it in this version of the TWES at the present time. Nevertheless, future empirical research may justify inclusion of a professional development subscale in subsequent versions. Validation is an ongoing process rather than a one-off event (Flora & Flake, 2017) and variations of the TWES should be systematically experimented with.

A third limitation relates to the external phase of this study. In particular, the outcome measures used are simple, single-item self-report measures. These short measures were used for pragmatic reasons related to limiting the length of the questionnaire and the resulting burden on respondents. However, these are unlikely to represent high-quality measures of the constructs of interest. With respect to turnover and attrition, we used self-reports of intentions of future job moves. However, turnover intentions have been shown to exceed observed turnover, which suggests our measure of turnover may be an overestimate (Ladd, 2011). Future research should therefore look to test whether the TWES predicts objective measures of turnover and attrition drawn from administrative data on the teaching workforce. With respect to job satisfaction, we used a simple single-item measure, which itself has not been validated. Future research should therefore look to test whether the TWES predicts validated measures of teacher job satisfaction (e.g. Thompson & Phua, 2012).

7.2 Future directions

The priority for future applied work must now be to collect individual-level panel data using the TWES. Collecting individual or subject/department level data will enable researchers to further exploit subject/department fixed effects, allowing them to account for additional unmeasured

potential confounders of the relationship between working environment and retention. Combining this data with information on changes in middle or senior leadership within schools would also provide an interesting opportunity to observe how working environment changes based on switches in leadership. Taken together, these approaches would allow analysts to get closer to understanding the underlying causal relationships between teachers' working environment, job satisfaction and retention.

A related priority for future research is to determine the relative importance of school-wide and department-specific working environment. For example, is a secondary school maths teacher's retention influenced more by the extent of supportive leadership provided by the Head of Maths or by the overall Headteacher of their school. The intra-cluster correlations reported in Table 4 suggest that teachers' perceptions of the quality of working environment are indeed more closely correlated within their department than across their school – which suggests that variation in working environment does indeed exist at the departmental level. However, further empirical research is required to assess the relative importance of the two levels for teacher retention and job satisfaction.

7.3 Conclusion

Many public schools systems face challenges in retaining teaching staff (Dolton, 2020; White & Smith, 2005), with high levels of turnover in specific schools reflecting difficult work environments. Research can help address this issue by identifying the aspects of teachers' working environment that are most strongly related to job satisfaction and retention, but only if working environment can first be measured accurately. The TWES (see Appendix Table 2 for the final scale) provides the first validated teacher self-report questionnaire for doing so. It can be used by researchers to systematically investigate which aspects of teachers' working environment are related to job satisfaction and retention or by school leaders looking to identify managerially controllable variables that they can improve in order to reduce staff shortage in their own schools.

References

Allen, R., Benhenda, A., Jerrim, J., & Sims, S. (2020). New evidence on teachers' working hours in England. An empirical analysis of four datasets. *Research Papers in Education*, 1-25.

Allensworth, E., Ponisciak, S., & Mazzeo, C. (2009). T*he schools teachers leave: Teacher mobility in Chicago Public Schools*. Chicago: University of Chicago.

Bakker, A. B., & Demerouti, E. (2017). Job demands–resources theory: taking stock and looking forward. *Journal of Occupational Health Psychology*, *22*(3), 273.

Barmby, P. (2006). Improving teacher recruitment and retention: the importance of workload and pupil behaviour. *Educational Research*, *48*(3), 247-265.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and practice*, *17*(1), 10-17.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061-1071.

Boyd, D., Grossman, P., Ing, M., Lankford, H., Loeb, S., & Wyckoff, J. (2011). The influence of school administrators on teacher retention decisions. *American Educational Research Journal*, *48*(2), 303-333.

Bozeman, B. (1993). A theory of government "red tape". *Journal of Public Administration Research and Theory*, *3*(3), 273-304.

Brunner, M., Nagy, G., & Wilhelm, O. (2012). A tutorial on hierarchically structured constructs. *Journal of Personality*, *80*(4), 796-846.

Bryk, A. S., Sebring, P. B., Allensworth, E., Luppescu, S., & Easton, J. Q. (2009). Survey measures, factors, composite variables, and items used in organizing schools for improvement: Lessons from Chicago. Retrieved from https://consortium.uchicago.edu/sites/default/files/2019-02/Survey-Development-Process-Appendix.pdf

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

Cameron, M. (2006). Managing school discipline and implications for school social workers: A review of the literature. *Children & Schools*, *28*(4), 219-227.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81.

Cheema, J. R. (2014). A review of missing data handling methods in education research. *Review of Educational Research*, *84*(4), 487-508.

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J. P., & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality*, *80*(1), 219-251.

Clifford, M., Menon, R., Gangi, T., Condon, C., & Hornung, K. (2012). *Measuring school climate for gauging principal performance: a review of the validity and reliability of publicly accessible measures*. Washington D.C: American Institutes for Research.

Collie, R. J., Shapka, J. D., Perry, N. E., & Martin, A. J. (2016). Teachers' psychological functioning in the workplace: Exploring the roles of contextual beliefs, need satisfaction, and personal characteristics. *Journal of Educational Psychology*, *108*(6), 788-799.

Cooper Gibson Research (2018). *Factors affecting teacher retention: qualitative investigation*. London, UK: Department for Education.

Crawford, E. R., LePine, J. A., & Rich, B. L. (2010). Linking job demands and resources to employee engagement and burnout: a theoretical extension and meta-analytic test. *Journal of Applied Psychology*, *95*(5), 834-848.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302.

Cuban, L. (1988). *The managerial imperative and the practice of leadership in schools*. Albany, NY: State University of New York Press.

DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: a literature review. *Quality & quantity*, *52*(4), 1523-1559.

Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry, 11*, 227–268.

Deci, E. L., & Ryan, R. M. (2008). Facilitating optimal motivation and psychological well-being across life's domains. *Canadian Psychology, 49*(1), 14-23.

Deci, E. L., Ryan, R. M., Gagné, M., Leone, D. R., Usunov, J., & Kornazheva, B. P. (2001). Need satisfaction, motivation, and well-being in the work organizations of a former eastern bloc country: A cross-cultural study of self-determination. *Personality and social psychology bulletin*, *27*(8), 930-942.

Demerouti, E., Bakker, A.B., Nachreiner, F., & Schaufeli, W.B. (2001). The job demands-resources model of burnout. *Journal of Applied Psychology, 86*, 499-512.

Dolton, P. (2020). Teacher supply. In Bradley, S., & Green, C. (Eds.), *The Economics of Education: A comprehensive overview* (2nd ed., pp. 391–402). Elsevier.

Donaldson, M. L., & Johnson, S. M. (2010). The price of misassignment: The role of teaching assignments in Teach for America teachers' exit from low-income schools and the teaching profession. *Educational Evaluation and Policy Analysis*, *32*(2), 299-323.

Donaldson, M. L. (2013). Principals' approaches to cultivating teacher effectiveness: Constraints and opportunities in hiring, assigning, evaluating, and developing teachers. *Educational Administration Quarterly*, *49*(5), 838-882.

Duyar, I., Gumus, S., & Sukru Bellibas, M. (2013). Multilevel analysis of teacher work attitudes: The influence of principal leadership and teacher collaboration. *International Journal of Educational Management, 27*(7), 700-719.

European Commission [EC] (2013). *Study on Policy Measures to improve the Attractiveness of the Teaching Profession in Europe: Volume 1*.

European Commission [EC] (2013). *Study on Policy Measures to improve the Attractiveness of the Teaching Profession in Europe: Volume 2*.

Fackler, S., & Malmberg, L. (2016). Teachers' self-efficacy in 14 OECD countries: Teacher, student group, school and leadership effects. *Teaching and Teacher Education, 56*, 185-195.

Fernet, C., Austin, S., Trépanier, S. G., & Dussault, M. (2013). How do job characteristics contribute to burnout? Exploring the distinct mediating roles of perceived autonomy, competence, and relatedness. *European Journal of Work and Organizational Psychology*, *22*(2), 123-137.

Fernet, C., Trépanier, S. G., Austin, S., Gagné, M., & Forest, J. (2015). Transformational leadership and optimal functioning at work: On the mediating role of employees' perceived job characteristics and motivation. *Work & Stress*, *29*(1), 11-31.

Fernet, C., Trépanier, S. G., Austin, S., & Levesque-Côté, J. (2016). Committed, inspiring, and healthy teachers: How do school environment and motivational factors facilitate optimal functioning at career start?. *Teaching and Teacher Education*, *59*, 481-491.

Feuerhahn, N., Bellingrath, S., & Kudielka, B. M. (2013). The interplay of matching and non-matching job demands and resources on emotional exhaustion among teachers. *Applied Psychology: Health and well-being*, *5*(2), 171-192.

Flake, J. K., & Fried, E. I. (2019). Measurement schmeasurement: Questionable measurement practices and how to avoid them. https://psyarxiv.com/hs7wm/download?format=pdf

Flake, J. K., Barron, K. E., Hulleman, C., McCoach, B. D., & Welsh, M. E. (2015). Measuring cost: The forgotten component of expectancy-value theory. *Contemporary Educational Psychology*, *41*, 232-244.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370-378.

Flora, D. B., & Flake, J. K. (2017). The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. *Canadian Journal of Behavioural Science*, *49*(2), 78-88.

Fowler, F. J. (2009). *Survey research methods* (4th ed.). Thousand Oaks, CA: Sage.

Gagné, M., & Deci, E. L. (2005). Self-determination theory and work motivation. *Journal of Organizational Behavior*, *26*(4), 331-362.

Gangi, T. A. (2009). School climate and faculty relationships: Choosing an effective assessment measure (Doctoral dissertation). *St. John's University, New York, NY*.

Gehlbach, H., & Brinkworth, M. E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. *Review of General Psychology*, *15*(4), 380-387.

Gibbons, S., Scrutinio, V., & Telhaj, S. (2018). *Teacher turnover: does it matter for pupil achievement? CEP Discussion Paper No. 1530*. London: LSE.

Grissom, J. A., Nicholson-Crotty, S., & Harrington, J. R. (2014). Estimating the effects of No Child Left Behind on teachers' work environments and job attitudes. *Educational Evaluation and Policy Analysis*, *36*(4), 417-436.

Gustafsson, J.-E., & Åberg-Bengtsson, L. (2010). Unidimensionality and interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (p. 97–121). American Psychological Association.

Gu, Q., & Day, C. (2013). Challenges to teacher resilience: Conditions count. *British Educational Research Journal*, *39*(1), 22-44.

Hakanen, J. J., Bakker, A .B., & Schaufeli, W. B. (2006). Burnout and work engagement among teachers. *Journal of School Psychology, 43*, 495-513.

Halbesleben, J. R. (2010). A meta-analysis of work engagement: Relationships with burnout, demands, resources, and consequences. *Work engagement: A handbook of essential theory and research*, *8*(1), 102-117

Hendricks, M. D. (2014). Does it pay to pay teachers more? Evidence from Texas. *Journal of Public Economics*, 109, 50-63.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, *66*(3), 393-416.

Hirsch, E., Freitas, C., Church, K., & Villar, A. (2008). *Massachusetts Teaching, Learning and Leading Survey: Creating school conditions where teachers stay and students thrive*. Retrieved from https://www.nysut.org/~/media/files/nysut/resources/2013/april/ted/mass_tlc_survey_finalreport.pdf?la=en

Hopkins, M., Bjorklund Jr, P., & Spillane, J. P. (2019). The social side of teacher turnover: Closeness and trust among general and special education teachers in the United States. *International Journal of Educational Research*, *98*, 292-302.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Ingersoll, R. M. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal*, *38*(3), 499-534.

Jarzabkowski, L. (2003). Teacher collegiality in a remote Australian school. *Journal of Research in Rural Education*, *18*(3), 139-144.

Jóhannesson, I. Á. (2006). 'Different Children—A Tougher Job'. Icelandic Teachers Reflect on Changes in Their Work. *European Educational Research Journal*, *5*(2), 140-151.

Johnson, S. M., & Birkeland, S. E. (2003). Pursuing a "sense of success": New teachers explain their career decisions. *American Educational Research Journal*, *40*(3), 581-617.

Johnson, S. M., Kardos, S. M., Kauffman, D., Liu, E., & Donaldson, M. L. (2004). The support gap: new teachers' early experiences in high-income and low-income schools. E*ducation Policy Analysis Archives*, *12*(61), 1-24.

Johnson, S. M., Kraft, M. A., & Papay, J. P. (2012). How context matters in high-need schools: The effects of teachers' working conditions on their professional satisfaction and their students' achievement. *Teachers College Record*, *114*(10), 1-39.

Kelchtermans, G. (2006). Teacher collaboration and collegiality as workplace conditions. A review. *Zeitschrift für Pädagogik*, *52*(2), 220-237.

Kelley, T. L. (1927). *Interpretation of educational measurements*. New York, NY: World Book Company.

Kraft, M. A., Marinell, W. H., & Shen-Wei Yee, D. (2016). School organizational contexts, teacher turnover, and student achievement: Evidence from panel data. *American Educational Research Journal*, *53*(5), 1411-1449.

Kraft, M. A., Papay, J. P., Johnson, S. M., Charner-Laird, M., Ng, M., & Reinhorn, S. (2015). Educating amid uncertainty: The organizational supports teachers need to serve students. *Educational Administration Quarterly*, *51*(5), 753-790.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50*, 537–567.

Ladd, H. F. (2007). Teacher labor markets in developed countries. *The Future of Children*, 201-217.

Ladd, H. F. (2011). Teachers' perceptions of their working conditions: How predictive of planned and actual teacher movement?. *Educational Evaluation and Policy Analysis*, *33*(2), 235-261.

Leiter, M. P., & Maslach, C. (1999). Six areas of worklife: A model of the organizational context of burnout. *Journal of Health and Human Resources Administration, 21*, 472–489.

Loeb, S., Darling-Hammond, L., & Luczak, J. (2013). How teaching conditions predict teacher turnover in California schools. *Peabody Journal of Education, 80*(3), 44-70.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635-694.

Lopes, J., & Oliveira, C. (2020). Teacher and school determinants of teacher job satisfaction: a multilevel analysis. *School Effectiveness and School Improvement, 31*(4), 641-659.

Marinell, W. H., & Coca, V. M. (2013). "Who stays and who leaves?" Findings from a three-part study of teacher turnover in NYC middle schools. Retrieved from https://steinhardt.nyu.edu/scmsAdmin/media/users/sg158/PDFs/ttp_synthesis/TTPSynthesis_Report_March2013.pdf

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary research and perspectives*, *15*(2), 51-69.

Meier, K. J., & O'Toole, L. J. (2010). Organizational performance: Measurement theory and an application: Or, common source bias, the Achilles heel of public management research. APSA 2010 Annual Meeting Paper.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, S69-S77.

Nathanson, L., Micha, S., Cole, R., Lent, J., McCormick, M. P., & Kemple, J. (2013). NYC School Survey 2008-2010: Descriptive report. *New York: Research Alliance for New York City Schools*.

National School Climate Council. (2007). The School Climate Challenge: Narrowing the gap between school climate research and school climate policy, practice guide- lines and teacher education policy. Retrieved from http://www.schoolclimate.org/ climate/advocacy.php

New Teacher Centre (2014). 2014 North Carolina Teacher Working Conditions Survey. Retrieved from https://ncteachingconditions.org/uploads/File/NC%20val%20rel%20brief%20%205-14.pdf

Organisation for Economic Co-operation and Development [OECD] (2010). *TALIS 2008 technical report*. Paris: OECD Publishing. Retrieved from https://www.oecd-ilibrary.org/education/talis-2008-technical-report_9789264079861-en

Organisation for Economic Co-operation and Development [OECD] (2014a). *TALIS 2013 technical report*. Paris: OECD Publishing. Retrieved from http://www.oecd.org/education/school/TALIS-technical-report-2013.pdf

Organisation for Economic Co-operation and Development [OECD] (2018). *PISA 2018 Results (Volume V): Effective Policies, Successful Schools*. Paris: OECD. See also: https://cutt.ly/Wkkirn1

Orphanos, S., & Orr, M. T. (2014). Learning leadership matters: The influence of innovative school leadership preparation on teachers' experiences and outcomes. *Educational Management Administration & Leadership*, *42*(5), 680-700.

Papay, J.P. & Kraft, M.A. (2017). Developing Workplaces Where Teachers Stay, Improve, and Succeed. In Quintero, E. (Editor). Teaching in Context: How Social Aspects of School and School Systems Shape Teachers' Development & Effectiveness. Harvard Education Press. 15-35.

Perryman, J., & Calvert, G. (2019). What motivates people to teach, and why do they leave? Accountability, performativity and teacher retention. *British Journal of Educational Studies*, 1-21.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, *63*, 539-569.

Prieto, L. L., Soria, M. S., Martínez, I. M., & Schaufeli, W. (2008). Extension of the Job Demands-Resources model in the prediction of burnout and engagement among teachers over time. *Psicothema*, *20*(3), 354-360.

Reeve, J., & Su, Y. L. (2014). Teacher motivation. In M. Gagne (Ed.), *The Oxford handbook of work engagement, motivation, and self-determination theory* (pp. 349-362).

Reeves, P. M., Pun, W. H., & Chung, K. S. (2017). Influence of teacher collaboration on job satisfaction and student achievement. *Teaching and Teacher Education*, *67*, 227-236.

Robinson, V. M., Lloyd, C. A., & Rowe, K. J. (2008). The impact of leadership on student outcomes: An analysis of the differential effects of leadership types. *Educational Administration Quarterly*, *44*(5), 635-674.

Ronfeldt, M., Loeb, S., & Wyckoff, J. (2013). How teacher turnover harms student achievement. *American Educational Research Journal*, *50*(1), 4-36.

Ryan, R. M., & Deci, E. L. (2011). A self-determination theory perspective on social, institutional, cultural, and economic supports for autonomy and their importance for well-being. In *Human autonomy in cross-cultural context* (pp. 45-64). Springer, Dordrecht.

Salokangas, M., Wermke, W., & Harvey, G. (2020). Teachers' autonomy deconstructed: Irish and Finnish teachers' perceptions of decision-making and control. *European Educational Research Journal*, *19*(4), 329-350.

Sebring, P. B., Allensworth, E., Bryk, A. S., Easton, J. Q., & Luppescu, S. (2006). The Essential Supports for School Improvement. Chicago: Consortium on Chicago School Research. Retrieved from https://consortium.uchicago.edu/sites/default/files/2018-10/EssentialSupports.pdf

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston: Houghton Mifflin,.

Shen, J. (1997). Teacher retention and attrition in public schools: Evidence from SASS '91. *Journal of Educational Research, 91*, 81–88.

Shrivastava, P., & Mitroff, I. I. (1984). Enhancing organizational research utilization: The role of decision makers' assumptions. *Academy of Management Review, 9*(1), 18-26.

Simkins, T. (2012). Understanding school leadership and management development in England: Retrospect and prospect. *Educational Management Administration & Leadership*, *40*(5), 621-640.

Simon, N. S., & Johnson, S. M. (2015). Teacher turnover in high-poverty schools: What we know and can do. *Teachers College Record*, *117*(3), 1-36.

Sims, S. (2018). *Essays on the recruitment and retention of teachers* (Doctoral dissertation). UCL.

Sims, S. (2020). Modelling the relationships between teacher working conditions, job satisfaction and workplace mobility. *British Educational Research Journal*, *46*(2), 301-320.

Skaalvik, E. M., & Skaalvik, S. (2011). Teacher job satisfaction and motivation to leave the teaching profession: Relations with school context, feeling of belonging, and emotional exhaustion. *Teaching and Teacher Education*, *27*(6), 1029-1038.

Skaalvik, E. M., & Skaalvik, S. (2017). Still motivated to teach? A study of school context variables, stress and job satisfaction among teachers in senior high school. *Social Psychology of Education*, *20*(1), 15-37.

Slemp, G. R., Kern, M. L., Patrick, K. J., & Ryan, R. M. (2018). Leader autonomy support in the workplace: A meta-analytic review. *Motivation and emotion*, *42*(5), 706-724.

Steen-Olsen, T., & Eikseth, A. G. (2010). The power of time: Teachers' working day—Negotiating autonomy and control. *European Educational Research Journal*, *9*(2), 284-295.

Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A. (2013). A review of school climate research. *Review of Educational Research*, *83*(3), 357-385.

Thompson, E. R., & Phua, F. T. (2012). A brief index of affective job satisfaction. *Group & Organization Management*, *37*(3), 275-307.

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements*. New York, NY: Columbia University Press.

Timms, C., Graham, D., & Cottrell, D. (2007). "I just want to teach" Queensland independent school teachers and their workload. *Journal of Educational Administration*, *45*(5), 569-586.

Toropova, A., Myrberg, E., & Johansson, S. (2021). Teacher job satisfaction: the importance of school working conditions and teacher characteristics. *Educational Review*, *73*(1), 71-97.

Torres, A. C. (2016). Is this work sustainable? Teacher turnover and perceptions of workload in charter management organizations. *Urban Education*, *51*(8), 891-914.

Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion*, *17*(2), 267.

Weiss, E. M. (1999). Perceived workplace conditions and first-year teachers' morale, career choice commitment, and planned retention: A secondary analysis. *Teaching and Teacher Education*, *15*(8), 861-879.

Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.
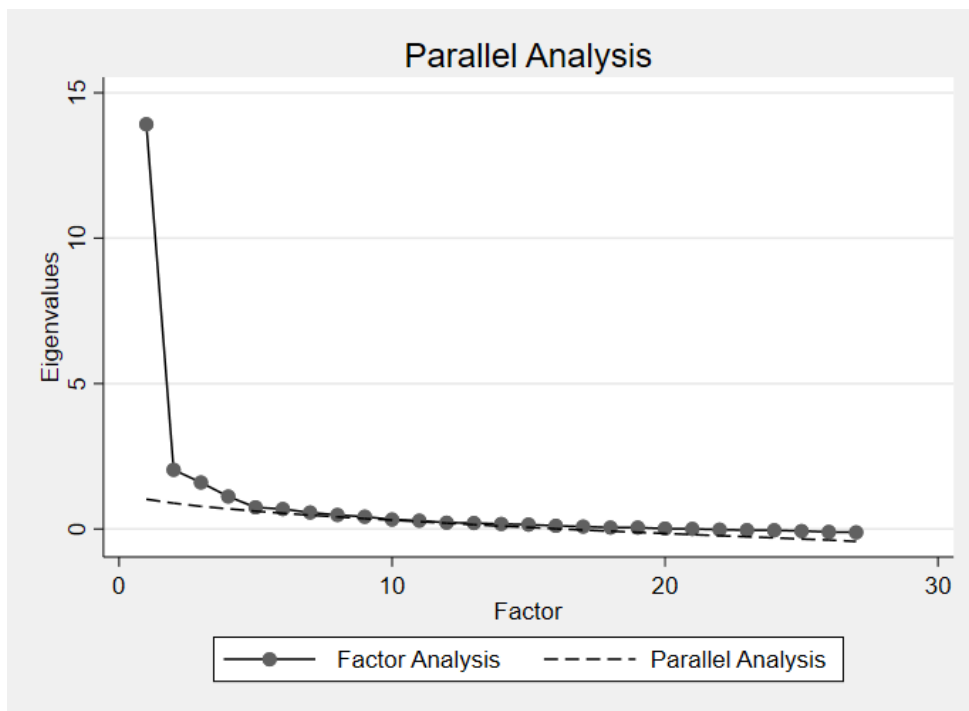
Figures

Figure 1. **Eigenvalues for the exploratory factor and parallel analyses (N=169)**

Tables

Table 1
Factor pattern matrix

	Item	Factor 1	Factor 2	Factor 3	Factor 4
Supportive Leadership	1	0.817			
	2	0.6507			
	3	0.8709			
	4	0.6475			
	5	0.6654			
	6	0.7447			
	7	0.6744			
	8	0.5061	0.4095		
Behaviour Policy	1		0.7832		
	2		0.77		
	3		0.9167		
	4		0.8852		
	5		0.872		
	6		0.761		
Collegiality	1				0.6634
	2				0.7612
	3				0.6334
	4				0.7594
	5				0.7731
	6	0.4148			0.5383
Compliance	1			0.6638	
	2			0.6178	
	3			0.5653	0.3415
	4			0.4736	
	5			0.8047	
	6			0.6668	

Note: After Promax rotation. Only loadings with an absolute value greater than 0.3 are shown. N=169.

Table 2

Reliability evidence

Factor	Item	Wording	IRC
Supportive Leadership α=0.94	1	Leaders say thank you to me for the work that I do	0.82
	2	Leaders are approachable	0.76
	3	Leaders recognise it when I do good work	0.80
	4	Leaders trust teachers to do the best for pupils	0.75
	5	Leaders can be relied upon for support if asked	0.82
	6	Leaders are willing to consider suggestions about new ways of doing things	0.76
	7	Leaders clearly communicate the reasons for which decisions have been made	0.76
	8	Leaders set a clear direction for school improvement	0.76
Behaviour Policy α=0.94	1	Expectations for behaviour are clearly communicated to pupils	0.78
	2	I can rely on middle & senior leaders to support me in relation to student behaviour	0.84
	3	The behaviour policy is properly enforced	0.88
	4	School policies help me manage classroom behaviour effectively	0.84
	5	I can rely on other staff to stick to the school behaviour policies	0.69
	6	The processes for dealing with behaviour are clear	0.84
Collegiality α=0.82	1	The teachers in my department/phase work together on planning	0.59
	2	The teachers in my department/phase are happy to help me out if I have a problem	0.69
	3	I feel comfortable asking colleagues for advice	0.54
	4	There are opportunities for me to learn from my colleagues	0.60
	5	I have the opportunity to use resources developed by my colleagues	0.65
	6	I have colleagues who care about me	0.58
Compliance α=0.81	1	I am expected to do things solely for the purpose of generating evidence	0.68
	2	I am expected to do tasks which do not contribute to pupils' education	0.63
	3	Requirements around lesson planning get in the way of teaching	0.52
	4	I am asked to teach subjects or age-groups for which I have not been trained	0.34
	5	Data management gets in the way of teaching	0.71
	6	Marking gets in the way of teaching	0.56

Note: IRC = Item-rest correlation within factor. α =Cronbach's alpha. N=169.

Table 3

Fit indices from the confirmatory factor analysis

	CFI	TLI	RMSEA	SRMR
1. One factor model	0.761	0.740	0.169	0.121
2. Three factor model	0.892	0.881	0.114	**0.073**
3. Four factor model	**0.955**	0.950	0.074	**0.047**
4. Five factor model	**0.956**	0.950	0.074	**0.045**
5. Higher order factor model	**0.959**	**0.955**	0.070	**0.048**
6. Nested factor model	**0.972**	**0.966**	0.061	**0.033**

Note: bold items show acceptable fit based on CFI>0.095; TLI>0.95; RMSEA<0.06; SRMR<0.08. N=1,231.

<div align="center">

Table 4

Pairwise and intra-cluster correlations

</div>

	SL	BP	CL	CP	School ICC	Dept ICC
Supportive Leadership (SL)	1.00				0.21	0.27
Behaviour Policy (BP)	0.64	1.00			0.34	0.39
Collegiality (CL)	0.47	0.33	1.00		0.05	0.24
Compliance (CP)	0.59	0.42	0.35	1.00	0.15	0.25

Note: N=1,231.

Table 5

Self-report: Ordered logistic regression analysis

	Job Satisfaction		Turnover Intentions		Attrition Intentions	
	(1) Separately	(2) Together	(3) Separately	(4) Together	(5) Separately	(6) Together
Leadership	5.08*	2.76*	0.32*	0.56*	0.44*	0.80*
	(0.39)	(0.29)	(0.02)	(0.05)	(0.03)	(0.08)
Behav. Pol.	2.95*	1.50*	0.41*	0.72*	0.51*	0.71*
	(0.18)	(0.29)	(0.02)	(0.07)	(0.03)	(0.06)
Collegiality	2.56*	1.56*	0.47*	0.69*	0.55*	0.74*
	(0.17)	(0.13)	(0.03)	(0.05)	(0.03)	(0.04)
Compliance	3.33*	1.84*	0.41*	0.66*	0.44*	0.60*
	(0.17)	(0.11)	(0.03)	(0.05)	(0.02)	(0.03)
IVs one-by-one	✓		✓		✓	
IVs together		✓		✓		✓
School FE	✓	✓	✓	✓	✓	✓
Pseudo R^2	-	0.14	-	0.16	-	0.11
N	1,230	1,230	1,228	1,228	1,231	1,231

Note. Columns (1), (3) and (5) each report the results of four separate ordered logistic regression models, in which each of the four independent variables (IVs) entered separately. Columns (2), (4) and (5) each report the results of one ordered logistic regression model, containing all four independent variables (IVs) entered together. Leadership = Supportive Leadership. Behav. Pol. = Behaviour Policy. School FE = school fixed effects. Standard errors clustered at the school level shown in parenthesis. * = $p < 0.05$.
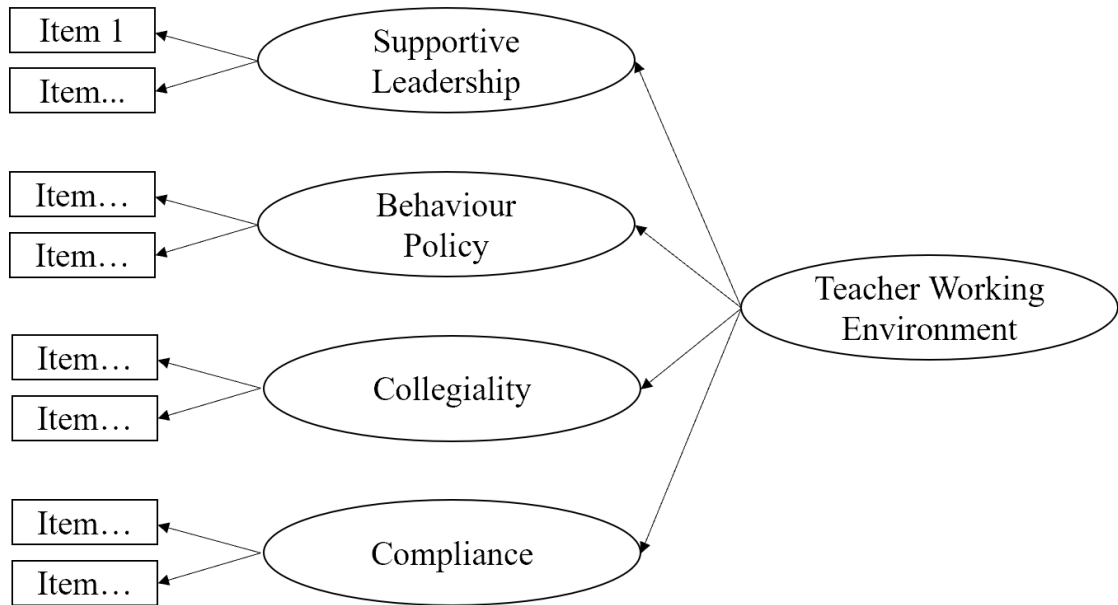
Table 6

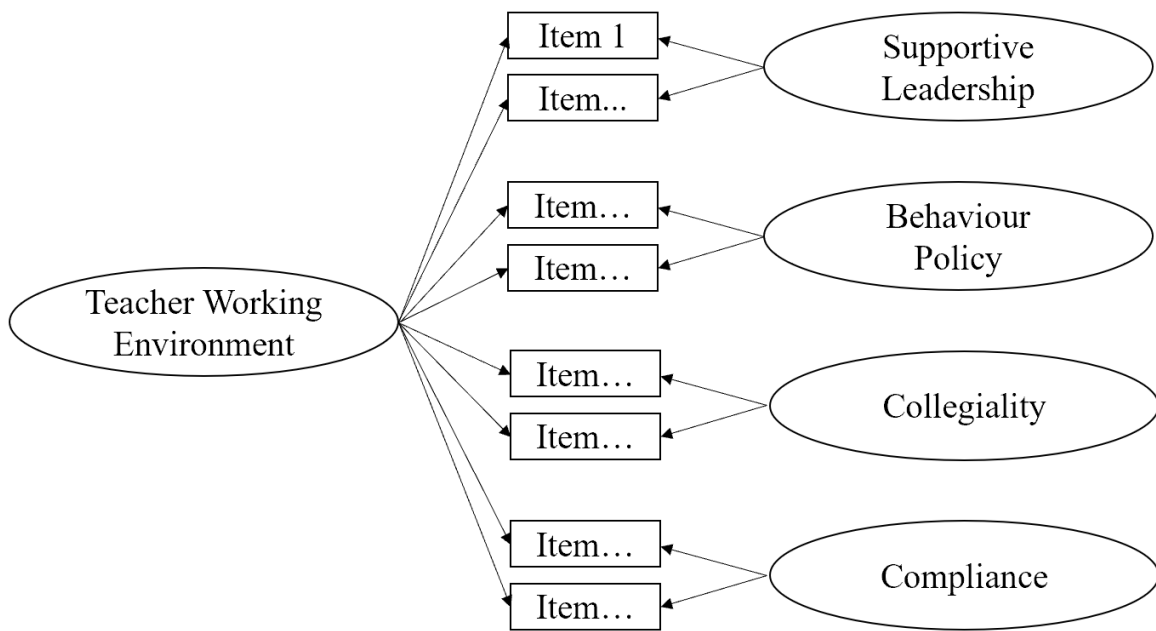**Colleague-report: ordered logistic regression analysis**

	Job Satisfaction		Turnover Intentions		Attrition Intentions	
	(1) All school colleagues	(2) Same subject colleagues	(3) All school colleagues	(4) Same subject colleagues	(5) All school colleagues	(6) Same subject colleagues
Leadership	2.45*	1.39*	0.52*	0.71*	0.61*	0.84
	(0.48)	(0.20)	(0.12)	(0.09)	(0.13)	(0.09)
Behav. Pol.	1.25*	1.33*	0.66*	0.67*	0.82	0.85
	(0.11)	(0.11)	(0.09)	(0.07)	(0.11)	(0.08)
Collegiality	0.76	1.13	1.06	0.80*	1.61*	0.87
	(0.18)	(0.98)	(0.26)	(0.05)	(0.35)	(0.07)
Compliance	1.27*	1.32*	0.98	0.93	0.71	0.76*
	(0.25)	(0.15)	(0.21)	(0.11)	(0.18)	(0.08)
Pseudo R^2	0.04	0.04	0.03	0.04	0.02	0.02
N	1,230	1,149	1,228	1,147	1,231	1,150

All columns report a single logistic regression in which all independent variables (IVs) are included. In columns (1), (3) and (5), the four working environment factor scores are calculated as the average score across all a focal teachers' colleagues within their school. In columns (2), (4) and (6), the four working environment factor scores are calculated as the average across all a focal teachers' colleagues who teach the same subject as the focal teachers. Leadership = Supportive Leadership. Behav. Pol. = Behaviour Policy. School FE = school fixed effects. Standard errors clustered at the school level shown in parenthesis. * = $p < 0.05$.

Appendix

Appendix Figure 1. **Schematic of the higher order factor model (row 5, Table 3)**

Appendix Figure 2. **Schematic of the nested factor model (row 6, Table 3)**