Psychometric Properties of Four Scales Measuring Teacher Wellbeing Evidence from Nigeria

Idalia Rodriguez Morales, Jorge Mahecha, Silvia Diazgranados Ferrans
International Rescue Committee

Technical Working Paper

October 2022

Abstract

Teachers are the most important school factor contributing to students' learning, wellbeing, and development (Jennings et al., 2017; Mendenhall, 2017; Renshaw et al., 2015) and as such, understanding teacher wellbeing (TWB) is critical to ensure a high-quality delivery of education. Reliable measurement with quality scales is a first step in promoting TWB by allowing to collect high-quality data that can enable to improve teacher educational interventions, policies, and programs. The present study aims to contribute to the limited pool of measures of TWB with evidence of validity and reliability in crisis and conflict-affected settings. Using data collected by the International Rescue Committee (IRC) in April and May of 2019 as part of a research study that aimed to evaluate the impact of a tutoring program for low-performing children in northeast Nigeria, we present the psychometric properties of four scales used to measure teachers' self-efficacy, job satisfaction, burnout, moods and feelings (focused on feelings of depression). For each scale, we analyze 1) reliability, 2) construct validity, 3) convergent and discriminant validity, and 4) concurrent validity. We complement the evidence on the psychometric properties of the four scales with recommendations for their revision, improvement, and contextualization for their use as measures in program evaluation and research.

Overview of Four Teacher Wellbeing Scales

The four scales analyzed in this study should have moderate to high evidence of validity and reliability for their use with evaluation purposes. The evidence in terms of validity is very good and in terms of reliability is fair. There is evidence of construct validity for individual scales. Between scales, there is fair evidence for convergent and construct validity, and good evidence of discriminant and face validity. Given the strength of the evidence, we recommend testing the psychometric properties of the scales with larger samples, test for inter-rater reliability, and consider revisions to the sales' items to ensure linguistic and contextual relevance, particularly in terms of adding items specific to the teaching profession. Given the guidance for revision and improvement for each tool, we recommend this measure for the specified purpose with adaptations and revisions specifically tailored to context.

Criteria	Indicators	Notes
Purpose	Program evaluation	All of the four scales require high internal consistency, and evidence of construct validity at the individual-scale level. Ideally, the different scales should correlate to each other following hypothesized patterns in order to provide evidence of construct validity for the higher-order construct of Teacher Well Being.
Empirical evidence overall	# of types of evidence available	7
	% Of evidence that meets empirical criteria	Between 54% (green only) to 96% (yellow and green) All scales meet criteria for internal consistency; 3 of scales meet criteria for discriminant validity. Evidence for construct validity is partial. All scales look face valid. Inter-rater reliability not assessed
Confidence in evidence	Sampling method	Convenience, Teacher-level: All teachers participating in an IRC program in Nigeria.
	Sample Size	Small (195, analytic samples varied between 149 and 179)
	Missing data	No Missing data, but several data anomalies (same scores for all teachers in a school/region, dropped from analyses)
Revisions	Clear guidance on what to adjust/refine	Yes: larger sample sizes with more attention to data integrity

Overview of Empirical Results (by tool)

Scale	Internal consistency	Face Validity	Construct validity	Convergent validity	Discriminant Validity	Concurrent validity	Recommendations /revisions
1. Self-Efficacy	✓	✓	○	☒	○	○	Improve sample sizes and data integrity. Consider revisiting items for cultural/ contextual relevance. Check for invariance
2. Job Satisfaction	✓	✓	✓	○	✓	○	
3. Burnout	✓	✓	○	○	✓	○	
4. Moods and Feelings	✓	✓	○	○	✓	✓	

Key

✓ Good/excellent evidence against empirical criteria	○ Fair/inconclusive evidence against empirical criteria	☒ Little to no evidence against empirical criteria	NA not applicable

This technical working paper was developed by Idalia Rodriguez Morales, Jorge Mahecha, and Silvia Diazgranados Ferrans for inclusion in the INEE Measurement Library using data collected by the International Rescue Committee (IRC) in Nigeria. We thank teachers and enumerators in Nigeria, IRC Nigeria country office staff, and IRC headquarters staff. We would like to acknowledge the support provided by Marlana Salmon-Letelier and Adane Miheretu.

Suggested citation: Rodriguez Morales, I., Mahecha, J., & Diazgranados Ferrans, S. (2022). *Psychometric Properties of Four Scales Measuring Teacher Wellbeing – Evidence from Nigeria.* Technical working paper. International Rescue Committee.

1. Introduction

Though there are different dimensions to wellbeing, it is generally defined as subjective life satisfaction in the emotional, psychological, and social aspects of life (Diener, 1984; Keyes, 2002). Teachers' role in education is instrumental, and as such, understanding teacher occupational wellbeing is critical to ensure a high-quality delivery in education. Although scarce, empirical evidence has pointed out to a positive relationship between teacher wellbeing (TWB) and student learning, safety and wellbeing (Mendenhall, 2017; Jennings et al., 2017), the quality of their instruction (Jennings 2016; McCallum, & Price, 2010), and to the negative relationship between TWB and teacher's turnover and absenteeism (Palma-Vasquez et al., 2022; Ryan et al., 2017; Falk et al., 2019). In crisis and conflict-affected settings, teachers face not only the usual challenges of impoverished education settings (crowded classrooms, limited resources, etc.) but they bear the effects of poverty, trauma, and conflict, alongside their students (Kirk & Winthrop, 2103; Wolf et al., 2015). Understanding TWB is thus critical to inform both programs and sector-level policies that support the delivery of high-quality education to improve student learning in the most needed settings.

The ability to generate high quality evidence on TWB, however, depends on the availability of high-quality measures and scales with evidence of validity and reliability; yet these are largely unavailable in contexts of conflict and crisis as much of the available evidence stems from the Global North (Diazgranados & Lee, 2020; Tubbs-Dolan & Yagoda, 2021). The aim of this study is to contribute to the availability of quality measures of TWB for use in conflict and crisis-affected settings. We present findings on the psychometric properties of four different scales of TWB that measure teachers' self-efficacy, job satisfaction, burnout, and moods and feelings (focused on feelings of depression). The data used for this study was collected by the International Rescue Committee (IRC) in April and May of 2019 as part of a research study that aimed to evaluate the impact of a tutoring program for low-performing children in northeast Nigeria.

2. Conceptual Framework

Teacher wellbeing (TWB) is a multidimensional construct, i.e. a construct that involves several different but associated facets (Hascher & Waber, 2021) and, as such, lacks an agreed-upon definition in the academic literature. This lack of a conventional definition is not a consequence of a lack of knowledge on the factors that determine TWB, but rather due to the abundance and diversity of individual, emotional, psychological, and social factors hypothesized to be related to TWB. Moreover, a recent review by the Education Equity Research Initiative (EERI) and the Inter-agency Network for Education in Emergencies (INEE) publication (Falk et al., 2019) further expands our understanding of TWB by using a socio-ecological framework to define TWB as a construct that encompasses teachers' affections, attitudes, and evaluations of their work across individual, school, community, and national, regional and international spheres.

Individual factors that affect TWB include gender, displacement status, employment status, level of education, teaching experience, coping mechanisms, content knowledge, and cultural competence (Falk et al., 2019). Emotional and psychological factors included in TWB conceptualizations include constructs of depression, stress, burnout, job satisfaction (Kim, 2019; Schleicher, 2018; Hall-Keyton et al., 2014; Cumming, 2017; Falk et al., 2019). Social and school-level factors affecting TWB include teacher quality and self-efficacy (Hall-Keyton et al., 2014; Falk et al., 2019; Capone & Petrillo, 2018), professional status, self-image and compensation (Kim, 2019; Schleicher, 2018), including teachers-student relationships, peer relationships such as coaching support and teacher collaboration (Hobson & Maxwell, 2017; Naghieh et al., 2015), school leadership, and school resources (Falk et al., 2019). Community, national and international factors include access to basic needs and policies around teacher management, compensation, certification, and right to work (Falk et al., 2019).

The measurement of TWB is complicated by the multidimensional nature of the construct and what they mean in various contexts. Research studies that indirectly conceptualize TWB use scales with teachers' samples that measure factors and/or constructs related to wellbeing, but without inquiring about the specifics of teaching or teachers. Studies in the Global South have used scales of a general nature related to health and/or other needs as proxies for wellbeing. For example, a study analyzing the psychological wellbeing of school teachers in India used the Short General Health Questionnaire (Thakur et al., 2018), a study with Arab teachers in Israel in Palestine used the Subjective Well-Being Assessment Scale, an instrument based on health science (Veronese et al., 2018), and a study of Palestinian teachers used the Global Well-being Assessment Scale (GWBAS) (Dakduki, 2015). In-depth psychometric evidence of validity and reliability of scales measuring TWB with an in-school or teacher-centric conceptualization originates mostly from the Global North[1], with limited evidence from the Global South, including the validation of the Wellbeing Holistic Assessment for Teachers (WHAT) tool in El Salvador (Soares et al., 2021), and the adaptation of the Teacher Subjective Wellbeing Questionnaire (TSWQ) with Brazilian educators (De Biagi et al., 2017).

A systematic review of 93 studies between 2000 and 2019 (Hascher & Waber, 2021) provides an updated overview of the field of TWB research, including aspects relevant to validity in measurement. The review suggests including not only positive but also negative components in the definition of TWB, for example, going beyond the absence of burnout or stress to highlighting the related positive and negative emotions. The study also points out that TWB research has been predominantly based on general concepts of wellbeing, rather than addressing the specific challenges and demands of the teaching profession.

[1] See, for example, Collie et al (2015) validate the Teacher Wellbeing Scale with a large sample of Canadian teachers; Laine et al. (2018) test the Occupational Well-being of School Staff Model (OWSS Model) in Finland and Estonia, Sadick & Issa (2017) test their Teacher Well-Being Survey in Canada; and Renshaw et al (2105) validate the Teacher Subjective Wellbeing Questionnaire (TSWQ) with elementary and middle school teachers in the USA and Mankin et al (2108) measure the same scales' invariance in the USA.

For this study, we use scales and data collected in Nigeria by the IRC in the context of a teacher's tutoring program. Data collection efforts included scales to measure TWB as IRC's Nigeria program and research teams deemed it a critical piece in the program's Theory of Change. Careful consideration was given to the selection of scales and items to balance an accurate reflection of the multifaceted nature of TWB in the context and alignment with previously validated tools.

The selection of scales and items included a review of the most salient factors affecting TWB, a scoping exercise of existing measurement tools, and a theoretical alignment and cross-cultural applicability and considerations. First, we conducted a comprehensive review of factors affecting TWB along with existing tools for their measurement, drawing from literature both from the Global North and Global South. This review provided an initial understanding of the array of factors related to TWB, from which country and program teams conducted a joint prioritization exercise and selected self-efficacy, job satisfaction, burnout, and depression as the more salient and appropriate for the context and study within the confines of the available resources and programmatic objectives. Scales focusing on these constructs were then cognitively pre-tested with a small sample of teachers in the study area, resulting in some deleted items deemed not necessary for the context and other semantic adjustments. Second, the theoretical underpinnings of each construct were carefully examined, with a particular focus on theories that offer insights into the dynamics of teacher well-being across different cultural contexts. This theoretical lens ensured that the constructs were not only universally relevant but also sensitive to the specific socio-cultural and educational challenges faced by teachers in the Global South and in our study context.

The theoretical framework guiding the scales' adaptation and combined use for the purpose of a holistic TWB measure, emphasizes the interaction between individual characteristics and factors with the broader socio-cultural environment. Drawing from Social Cognitive Theory (Bandura, 1977), which posits that individuals' beliefs in their capabilities to exert control over their actions and events that affect their lives, we hypothesize that teacher self-efficacy, defined as a teacher's belief that s/he is able to guide and elicit desired outcomes for student behavior, learning, and achievement (Skaalvik & Skaalvik, 2007; Falk et al., 2019), is associated with greater enthusiasm, commitment, and resilience, which in turn enhances overall wellbeing and job satisfaction defined as the sense of fulfillment and gratification receiving from his or her occupation (Hall-Kenyon et al., 2014; OECD, 2020). We base our hypothesis on studies in fact suggest that teachers with high self-efficacy are more likely to exhibit resilience, and experience greater job satisfaction, which collectively contribute to enhanced wellbeing.

We conceptualize burnout, as a type of work-related stress that can manifest as exhaustion and a reduced sense of accomplishment (Bermejo-Toro et al., 2016; Buric et al., 2019), and negative feelings related to depressive symptoms, such as lack of energy, loss of interest and feelings of frustration and alienation (Capone & Petrillo, 2018). Given the challenging working conditions of teachers in Northeast Nigeria, including resource scarcity, political instability, and socioeconomic pressures, teachers we hypothesize they might be particularly susceptible to burnout. This condition could be exacerbated by the high expectations placed on teachers by communities and the

emotional toll of teaching in such environments, causing burnout to be more pronounced where there is a significant disparity between the needs of the students and the resources available to teachers, potentially negatively impacting wellbeing. However, we also hypothesize that the positive self-assessment stemming high levels of self-efficacy can mitigate feelings of burnout by promoting a sense of personal accomplishment and control over work-related stressors. We focus on burnout and depression as separate but interrelated constructs, and do not include anxiety, first, in terms of measurement economy, length of the survey and wellbeing of respondent, and second, due to the comorbidity between anxiety and depression (Ter Meulen et al., 2021) and the measurement of both could introduce variability and complexity in the interpretation of results. Addressing depression exclusively does not negate the importance of anxiety or the comorbidity between the two; rather, it provides a clear, manageable scope for the current study while acknowledging the exploration of comorbidity as an important direction for future research.

Lastly, from the Job Demands-Resources (Bakker & Demerouti, 2007; Granziera et al. 2021) Model, we draw that high job demands can lead to burnout, whereas access to job resources, including social support and constructive feedback, can foster job satisfaction and buffer against the negative impacts of job demands on mental health. This model helps spell out the dynamic interplay between external job factors and internal psychological states, illustrating how they collectively influence teacher wellbeing by either exacerbating or alleviating stress, burnout, and depression. Together, these theoretical perspectives provide a multifaceted understanding of the factors contributing to teacher well-being in the Global South, highlighting the complex interactions between individual psychological traits, job characteristics, and broader environmental factors.

For our study's purpose and based on the available scales' data prioritized by program teams, we thus conceptualize TWB as encompassing high levels of self-efficacy and job satisfaction and low levels of depressive factors and burnout. Finally, complementing our analysis with the relates scales exploring teachers' perceptions and experiences with coaching, in-schools social support and collaboration, and teaching barriers, we emphasize the importance of agency, environmental resources, and wellbeing. Based on this theoretical framework, we consider the combination of these scales in one measure suits the purpose of exploring TWB in the Global South. The used scales also fit INEE's TWB review and conceptualization, with scales addressing wellbeing both in a general way (job satisfaction), and directly inquiring about the teaching profession (self-efficacy and burnout). Second, the scales also address the suggestions on measurement from Hascher and Waber (2021) of including both positive – e.g., job satisfaction and self-efficacy– and negative – e.g., burnout, feelings of anxiety and depression–related wellbeing aspects.[2] Moreover, because the program also collected additional data specifically related to teaching challenges and activities (such as coaching, in-school social support, collaboration and relationships with peers, and other barriers to teaching), the scales and data provide an opportunity to analyze TWB in direct relationship with challenges experienced by teachers in crisis and conflict-affected settings.

[2] The Moods and Feelings scale is intended to be a proxy for depression and, while not directly related to occupation or job stress it is a helpful scale to measure emotional factors of wellbeing.

3. The Present Study

Using the data collected in Nigeria, this study first presents evidence of validity and reliability of scales measuring individual-level constructs of TWB (self-efficacy, job satisfaction, burnout, and moods and feelings). The study then delves into correlations between TWB and contextual factors that the literature identifies as being either related or integral to TWB (coaching experience, social support and collaboration, and barriers to teaching) to provide evidence of convergent, discriminant, and concurrent validity.

3.1 Research questions

The present study aims to assess the psychometric properties and evidence of validity and reliability of different TWB scales in a context of conflict and crisis. In particular, we aim to answer the following questions:

1. How do the different scales capture variations in TWB dimensions? What scale behavior and response patterns do we observe?
2. Is there evidence of reliability?
3. Is there evidence of construct validity?
4. Is there evidence of convergent and discriminant validity?
5. Is there evidence of concurrent validity?

3.2 Data

The data used in this study was collected by the IRC in April and May of 2019 as part of a study about a tutoring program for low-performing children in northeast Nigeria. The program was implemented by government school teachers that worked as tutors in an afterschool program in northeast Nigeria. Enumerators were trained in data collection procedures and ethics in a 3-day collective training. For the TWB data, enumerators assisted in respondents with reading the questions and recording responses. TWB data consisted of 195 registries, one registry per teacher with a unique identification code and date and time stamps of data collection. We calculated the average length of the survey to identify anomalies in data collection, i.e., excessively short times for survey completion. The complete set of six scales plus the background information of teachers (grades they teach, years of experience, gender, etc.) involves 104 prompts. If each prompt takes up to 2 seconds to read and respond to, it can be assumed that survey responders taking less than 208 seconds (104 x 2 = 208), probably responded to survey questions automatically without actually engaging in reading the prompts. This resulted in the deletion of 12 teacher registries. Two additional registries that provided no response to any of the scales were deleted as well. This resulted in a sample of 181 teachers corresponding to primary school teachers (82 female and 99 male) in two Nigerian states: Borno and Yobe, and a total of 10 communities (7 and 3 communities, respectively). All survey tools were implemented in English, being that all respondents were fluent in English.

Table 1 below provides an overview of the teachers in the sample and their characteristics. Teachers belong to 80 different schools, with an average of 18 teachers surveyed per school. The teachers range in age from 22 to 58 years old (mean=41.5, SD=8.7). Most of the teachers (74%) hold a Nigerian national Certificate of Education (NCE), a 3-year college certificate – the minimum qualification required to teach in primary and junior high schools–, 18% of the teachers hold a Bachelor's degree, and the remaining 8% of the sample are distributed among teachers with a National Certificate, with some graduate studies, and those who only completed secondary school.

Table 1. Teachers and schools by states and communities

State and Community	N (schools)	N (teachers)	% female teachers	Teachers' mean age
Borno	*58*	*154*	*47*	*41.7*
Biu	9	32	47	45.8
Gwoza	4	18	22	42.3
Jere	8	20	50	38.2
Kaga	5	11	36	43.1
Konduga	8	15	33	43.2
MMC	21	49	60	39.3
MONGUNO	3	9	56	42.1
Yobe	*22*	*27*	*37*	*40.7*
Bade	3	3	33	38
Damaturu	13	14	36	38.4
Potiskum	6	10	40	44.5
Overall	80	181	45	41.5

After examining patterns of responses by geographic location, some data anomalies were detected in the state of Borno, specifically in the community of Biu, for which the scales had no variation (SD=0), i.e., all teachers in a particular subset of schools with the same score: a) for eight schools (28 teachers) there was no variation (SD=0) for the scale teacher efficacy scale, b) for nine schools (32 teachers) there was no variation for the Burnout scale, c) for four schools (15 teachers) there was no variation for the Job Stress), and d) for eight schools (29 teachers) all teachers had the same score for the Coaching Experience scale. Other communities in the state of Borno showed similar anomalies but involved fewer schools and teachers. Because it is highly unlikely that teachers in the same schools and communities would have both identical scores and the exact same response patterns, this anomalous data was excluded from psychometric analyses in this study in the analyses of each corresponding scale.

3.3 Scales

We used four scales to measure different aspects of teacher wellbeing, specifically: Teacher Self-Efficacy, and Job Satisfaction scales, which address positive aspects of TWB, and the Moods and

Feelings and Burnout scales, which address negative aspects of TWB. In this section, we detail the TWB and associated scales collected by the program.

3.3.1 Self-Efficacy

The teacher self-efficacy scale used in this investigation includes 10 items using a five-point Likert scale asking respondents to indicate the level of agreement (“Not at all”, “Very little”, “Moderate”, “Quite a bit” and “A great deal”). This is an ad-hoc scale that bears some similarity to previous scales that measure this construct, such as Bandura’s teaching self-efficacy scale (Bandura, 2006) or the Norwegian Teacher Self-Efficacy Scale (Skaalvik & Skaalvik, 2007), touching on themes that these scales cover, like the ability of the teacher to effectively manage the classroom or motivate students. Items cover various aspects of teaching, including parental engagement, class management and motivation (e.g., “I can motivate students who show low interest in school”; “If a student in my class is undisciplined, I know some techniques to redirect him or her”). The scales’ maximum score is 50 points, and there are no reverse-coded items. For all items, higher scores indicate more of the attribute.

3.3.2 Job-Satisfaction

The scale for job satisfaction is a short (4-item) scale using a six-point Likert scale of agreement (“Strongly disagree”, “Disagree”, “Moderately disagree”, “Moderately agree”, “Agree”, and “Strongly agree”). The items included in the scale do not exactly align with those of other scales measuring job satisfaction in teachers (Gkolia et al., 2014; Ho & Au, 2006; Skaalvik & Skaalvik, 2011, 2014) but bear enough similarities to be based on those. At face value, the items are of general nature and are not specifically related to aspects related to teaching but to satisfaction with the individual’s current job, such as “I want to continue for a long time in my current workplace” or “I enjoy being in my current job position”. This scale's maximum score is 24 points, and no items are reverse-coded. For all items, higher score points indicate more of the attribute.

3.3.3 Moods and Feelings

The ad-hoc 18-item scale, which reflects an adaptation of the Goldberg Anxiety and Depression Inventory (Goldberg et al, 1998), which uses a five-point Likert scale inquiring for frequency (“Never”, “Rarely”, “Sometimes”, “Very often”, and “Always”) of psychological and physical manifestations of depression such as “Felt that you have lost weight (due to poor appetite)?” or “Felt that you have lost confidence in yourself?”. The maximum scale score is 90 points, and no items are reverse-coded. For all items, a higher score points indicate more of the attribute, i.e., more stress.

3.3.4 Burnout

The burnout scale used in this report was first suggested by Richmond and Gorham (1992). It is a 20-item scale using a five-point Likert scale inquiring about frequency of behaviors ("Not at all", "Very little", "Moderate", "Quite a bit", "A great deal"). Items point to negative perceptions associated with work such as "I feel sick to my stomach when I think about work", or more specific perceptions related to working in schools such as "I dread going into a classroom". None of the items are reverse-coded. This scale's maximum score is 100 points, with higher scores corresponding to more burnout.

3.3.5 Associated scales

In addition to the TWB scales described before, three additional scales, Coaching Experience, Social Supports and Collaboration, and Teaching Barriers are included in this report as they measure constructs associated with TWB. The coaching experience scale is a 12-item scale using a five-point Likert scale ("Not at all", "Very little", "Moderate", "Quite a bit", "A great deal") to ask about the frequency of certain behaviors displayed by teacher's coaches, for example, "To what extent does your coach let you know what she/he expects from you"; "… guide you in a right direction' or "…talk to you about your behavior".

The Social Support and Collaboration scale is a short (5-item) scale that inquires about the frequency of behaviors related to collaboration between teachers in tasks such as discussing lesson plans with colleagues, observing colleagues, or working together. It uses a four-point Likert scale ("Never or almost never", "2 – 3 times per month", "1 – 3 times per week", "Daily or almost daily"), does not have any items that require reverse-coding, and higher score points indicate more of the attribute.

The Barriers to Teaching scale is a six-item scale describing a series of factors that could make difficult for a teacher to teach students, like the lack of foundational knowledge by students, behavioral issues or contextual factors affecting students like lack of proper feeding, lack of sleep, etc.

4. Analytic approach and hypotheses

For research question 1, we analyzed the ranges and distribution of each scale using descriptive analyses. To answer research question 2 on scales' reliability, we calculated Cronbach's alpha for each scale, item-to-scale correlations, and the change in reliability if an item is removed from each instrument scale. Cronbach's alpha is a coefficient for measuring a scale's reliability, i.e., the degree to which the items can be said to measure the same concept or construct in a consistent way (Tavakol, 2018). It varies between 0 and 1, with values equal to or above 0.7 being considered conventionally acceptable for research purposes (Bujang et al., 2018).

To answer research question 3 on the scales' construct validity, i.e., "the extent to which the scores obtained using a particular measuring instrument agree with theoretical expectations" (Knapp, 2011), we conducted exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) for each scale to confirm if the hypothesized latent constructs of the measures were consistent with the data. EFA is a statistical method used to analyze patterns in the joint variation of items in a scale or test. If some items are found to jointly vary, this could be the basis for stating that the analyzed items constitute a factor, i.e., a hypothesized construct within a scale (DeVellis, 2012). In the absence of substantial theory about the structure of correlations between a scales' items and constructs, which is often the case in new or ad hoc scales, EFA is the option of choice (Fabrigar & Kan, 2018). CFA is a statistical method that uses analysis of covariance to confirm hypothesized relationships among a set of variables. Unlike EFA, CFA involves a preexisting theory about the factors and their relationships on a scale (Randall & Hung, 2018). In our review, we could not find literature describing measurement models for any of the scales used for this study. Hence, we decided to use existing literature relevant to the constructs included in each scale to first run an EFA which, along with substantive considerations of the content of the scale provided the theoretical expectation on which to run a CFA.

To conduct the preliminary EFA exploration, a single factor EFA analysis was run on all scales to obtain the number of eigenvalues larger than one, and the variance explained to get a sense of the possible number of factors involved. When these criteria indicated that more than one factor was necessary, the sample was divided into an EFA sample and a CFA sample. Following best practices and guidance, we randomly chose and used approximately 40% of the available sample for the EFA model and the remaining 60% for the CFA model. Although the sample for this study is relatively small for EFA or CFA according to traditional thresholds (Hair, 2010), simulation studies have proved that under conditions of a small number of factors, high loadings values, and no less than 6 items per scale, it is possible to perform EFA with samples below 50 (de Winter et al., 2009), and CFA with samples between 60 and 90 (Mundfrom et al., 2005; Wolf et al., 2013).

Then, based on the literature reviewed, we used CFA to test the following hypotheses:

- Teacher self-efficacy has a two-factor latent structure based on the fact that items in the scale could be classified into two broad categories: items related to teacher skills (3, 4, 5, 7, 8) and items related to class management (1, 2, 6, 9, 10). Other scales for Teaching Self-Efficacy have used a similar classification of items (APA, 2020; Bandura, 2006; Skaalvik & Skaalvik, 2007)
- The four items of the Job Satisfaction scale conform to a single factor.
- The Moods and Feelings scale has a two-factor structure related to behavioral sources of stress (items 2-5, 10-13, 17) and physical sources of feelings of depression (items 1, 6 – 9, 14 -16 and 18).
- The Burnout scale loads into two factors: general sources of burnout (items 1, 3-7, 9-11, 15, 16, 18-20) and school-related sources of burnout (items 2, 8, 12-14, 17).

For all the analysis, we used fit indices (CFI, TLI, RMSEA) to evaluate goodness of fit. Additionally, we examined item loadings and R-square of each item. Factor loadings reflect the degree to which

each item is linked to the hypothesized factor. If an item is hypothesized to load on a particular factor, then its loading should be large, positive, and statistically significant, with values larger than .50 being desirable. The R-square measures the proportion of variance in Y that is explained by the model and will be between 0 and 1, with larger values indicating a higher proportion of variance in the endogenous variable explained by the model.

To answer research question 4 on convergent and discriminant validity, i.e., the degree to which the score of an instrument predicts an outcome based on information from external instruments (Kline, 2011), we first estimated sum scores for the scales and individuals in the sample and then estimated correlations between the sum scores of the different scales to check if they follow hypothesized patterns of size and direction. For our study, we use thresholds for weak, moderate, and strong or high Pearson correlation coefficients as |0.3|, |0.5|, and |0.7|, respectively, where the vertical lines around a number denote its absolute value (i.e., its magnitude disregarding its sign) in accordance with recent research and revisions on sizes of correlation (Courtney, 2018; Mukaka, 2012). According to the literature, we hypothesized the following correlations among the TWB scales and the associated scales (See ***Table 2***).

Table 2. Predicted correlations between TWB scales and associated constructs

Scale	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1. Self-Efficacy	1						
2. Job Satisfaction	Strong, positive	1					
3. Burnout	Strong, negative	Strong, negative	1				
4. Moods and Feelings	Moderate, negative	Negative	Strong, positive	1			
5. Coaching Experience	Weak, not negative	Weak	Moderate, negative	Weak, not positive	1		
6. Social Support and Collaboration	Weak, not negative	Weak, positive	Moderate, negative	Weak, not positive	Weak	1	
7. Teaching Barriers	Moderate, negative	Moderate, negative	Moderate, positive	Moderate, positive	Weak	Weak, positive	1

1. Teacher Efficacy scale

- Strong and positive correlation with Job Satisfaction: It seems unlikely that an individual with high self-efficacy in teaching, characterized by success in achieving learning with students, does not have a high level of job-satisfaction.
- Strong and negative correlation with Burnout: A high score on the teacher efficacy scale is indicative of good relationships with students and a positive view of class interactions. The teacher burnout scale, on the other hand, assigns high scores to self-reported feelings of an antagonistic and troubled relationship with students.

- Moderate and negative correlation with Moods and Feelings. A high score on the Moods and feelings scale indicates a high prevalence of negative feelings related to job stress, which could translate into a self-perception of low efficacy.
- Weak and not negative correlation with coaching experience: If teachers have undergone coaching and their experience was positive, this might be positively associated with self-efficacy, though not necessarily. However, teachers might have high self-efficacy without having undergone any coaching.
- Weak and not negative correlation Social Support and Collaboration: A high score on the Social Support and Collaboration scale indicates frequent, constructive interactions with peers and colleagues. While it could be the case of a positive correlation between self-efficacy and these interactions, it is not necessarily the case that someone with high self-efficacy has to have these kinds of interactions on a frequent basis.
- Moderate and negative correlation with teaching barriers: A teacher with high self-efficacy would most likely not have a high score on the teaching barriers scale, as he or she declares being able to overcome different obstacles to learning.

2. Job Satisfaction scale

- Strong and negative correlation with Burnout: These concepts are conceptual opposites, i.e., it is unlikely to be highly satisfied with a job that causes high levels of burnout.
- Negative correlation with Moods and feelings: Feelings of depression and anxiety could impact job satisfaction.
- Weak correlation with Coaching Experience: Coaching experience does not necessarily imply anything about job satisfaction; we hypothesize a weak correlation in either direction.
- Weak and positive correlation with Social Support and Collaboration: Coaching does not necessarily involve the group activities implied in the Social Support scale.
- Moderate and negative correlation with Teaching Barriers: A teacher scoring high on Teaching Barriers is likely to be frustrated with the difficulties of teaching and hence, show low levels of job satisfaction.

3. Burnout scale

- Strong and positive correlation with Moods and Feelings: A high score on the Moods and feelings scale indicates the persistence of negative feelings of anxiety and depression, which can also relate to feelings of burnout.
- Moderate and negative correlation with Coaching Experience: Under the assumption that coaching has positive effects and improves how the teacher feels at work, however, we hypothesize a moderate correlation since coaching does not necessarily address aspects related to burnout.
- Moderate and negative correlation with Social Support and Collaboration: A high score in the Social Support and Collaboration indicates frequent positive interactions with peers, in direct

contrary to items in the Burnout scale such as “I dislike going to my job”, or “I avoid communication with my colleagues”.

- Moderate and positive correlation with Teaching Barriers: While someone experiencing burnout might also be experiencing an increased number of barriers to teaching, burnout can also have intrinsic components, so the relationship between these two scales could be moderate.

4. Moods and Feelings scale

- Weak correlation with Coaching Experience: Feelings of anxiety and depression do not necessarily imply positive or negative feelings around teachers’ coaching experience.
- Weak correlation with Social Support and Collaboration: Feelings of anxiety and depression do not imply a relationship in either direction on the social supports scale
- Moderate and positive correlation with Teaching Barriers: High scores in the Moods and Feelings scale could make someone more susceptible to overstate the degree to which they experience teaching barriers but could also have other causes.

5. Coaching Experiences scale

- Weak correlation with Social Support and Collaboration: Coaching experience and social support are different constructs with no apparent directional relationship between them.
- Weak correlation with Teaching Barriers: Receiving high-quality coaching could not have a strong relationship with teachers’ perception of the teaching barriers they experience.

6. Social Support and Collaboration

- Weak correlation with Teaching Barriers: A high frequency of interaction with teacher peers might help in addressing the perception of the magnitude of teaching barriers or could work as a resonance box exacerbating them. Hence, the correlation could go either way, and it is not predicted to be strong or moderate.

We evaluated the scales using goodness-of-fit indicators, which include the Comparative Fit Index (CFI), root mean square error of approximation (RMSEA), and Standardized Root Mean Square Residual (SMRS). The CFI is an incremental fit index that compares the theoretical model with a null model and uses the non-central chi-square distribution. A CFI greater than .95 indicates a good fit of the model. The RMSEA corresponds to the root mean square error of approximation; an RMSEA of .05 is considered to demonstrate optimal fit of the model. The SMRS is a measure of the mean absolute value of the correlation residuals; an SMRS smaller than .05 indicates good fit (Kline, 2011).

Finally, to answer research question 5 on concurrent validity, which implies comparing one established scale measuring a certain attribute or construct with another scale measuring a related attribute (West & Beckman, 2018), we compared the TWB scales with data from three additional scales: Coaching Experiences, Social Support and Collaboration, and Teaching Barriers.

5. Results

5.1 What scale behavior and response patterns do we observe?

Following the original tools' scoring procedures, a total score was defined for each of the scales as the simple sum of the coded values for the items. For example, for the teacher self-efficacy scale, which has 10 questions, the maximum possible score is 50, because the scale point indicating the highest level of this trait for each of the items ("A great deal") was coded as a five (10 questions x 5 points = 50 points). A maximum score was defined similarly for each scale. Higher scores always indicated more of the trait, as indicated in the methods section: higher scores in burnout indicate more burnout; higher scores in job satisfaction indicated more job satisfaction, etc.

The distributions for all the scales are in ***Figure 1***, and basic descriptive statistics for each scale (minimum and maximum scores, average, and standard deviation) are in Table 3. Only data corresponding to registries that were not considered anomalous are included in all analyses (see section 3.2, Data).

Table 3. Scales - Descriptive statistics

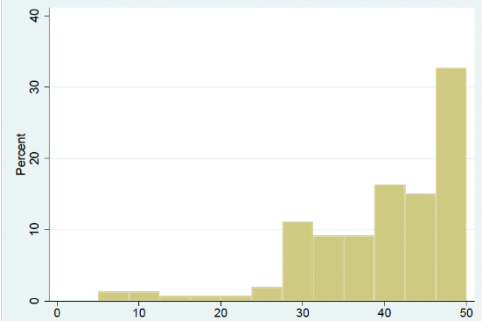
Scale	Obs	Mean	Std. Dev	Min	Max
Self-Efficacy	153	40.2	9.5	5	50
Job Satisfaction	179	19.7	5.6	4	24
Moods and Feelings	166	29.2	12.6	1	76
Burnout	149	26.0	12.1	1	76
Coaching experience	152	47.3	12.1	5	60
Social Support and Collaboration	175	15.1	4.2	2	20
Teaching Barriers	179	17.8	4.8	9	25

Results indicate high scores in the Self-Efficacy and in the Job satisfaction scale, with close to 80% of the responses at the top of both scales; and low scores in the Burnout – close to the scale's minimum– and the Moods and Feelings scales – with a mean score below 40% of the scale's maximum score. The scales for Coaching Experience, Social Support and Collaboration, and Teaching Barriers show high scores relative to maximum scale scores, indicating positive experiences with coaching, high levels and quality of interaction with peers, and high levels of teaching barriers, respectively.

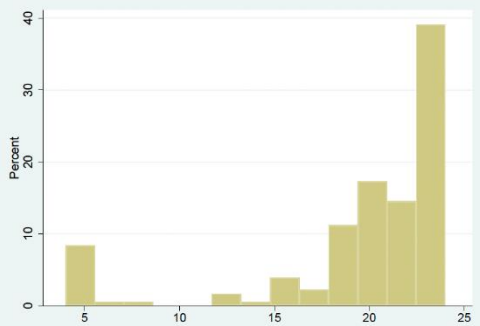
Figure 1. Distribution of all scales

A. Self-Efficacy

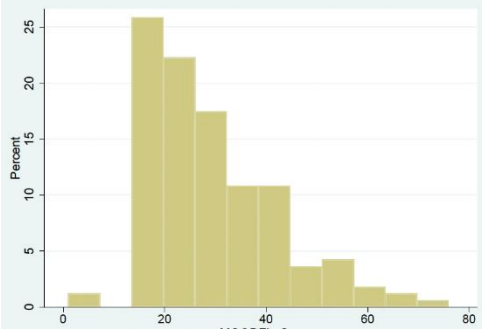
B. Job Satisfaction



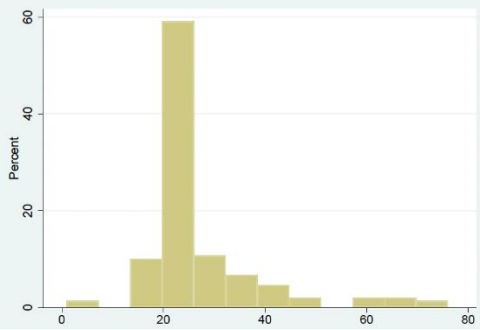
C. Moods and feelings



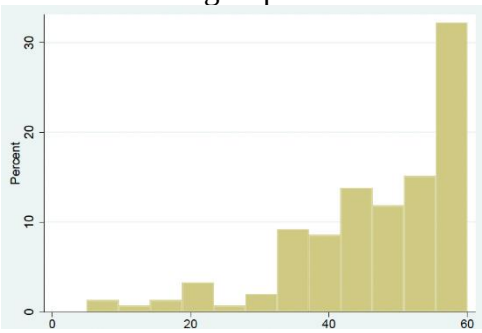
D. Burnout

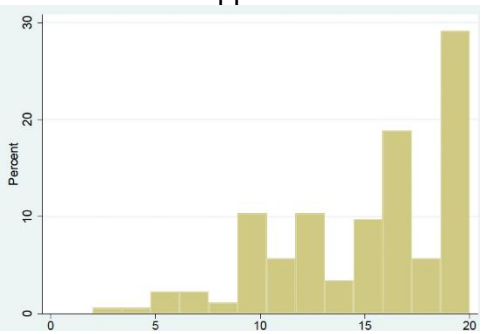


E. Coaching Experience

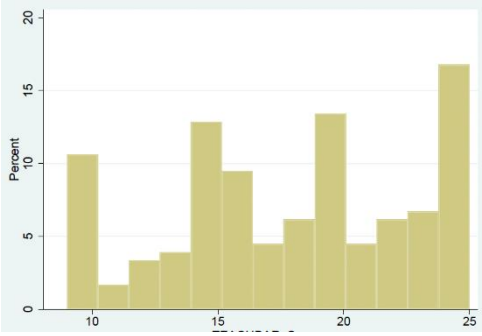


F. Social Support and Collaboration





G. Teaching Barriers



While overall results show a positive picture for teachers, when looking at the community level, some noteworthy differences become visible. Table 4 includes color coding highlights in red values that indicate lower wellbeing. For instance, the community of Jere (8 schools, 20 teachers) shows the lowest self-efficacy average score of all communities, almost 10 points below the average for the scale. The community of Monguno (3 schools and 9 teachers) shows a concerning pattern of highest scores for burnout, job-related stress (Moods and feelings), low job satisfaction. Although the differences between schools and communities might not be current, their presence highlights the importance of disaggregating the data at the community and even the school level.

Table 4. Scales scores by state and community

State	Community	Self-Efficacy	Burnout	Job Stress	Job Satisfaction	Coaching experience	Social Supports and Collaboration	Teaching Barriers
Borno	Biu	49.72	20.00	18.64	16.32	59.37	16.92	13.51
	Gwoza	47.15	25.89	31.19	20.83	51.45	14.90	21.67
	Jere	32.69	19.83	24.04	22.25	48.92	11.27	18.54
	Kaga	47.10	22.07	22.13	22.17	55.23	18.60	13.23
	Konduga	34.00	26.15	32.75	20.15	44.94	15.19	17.08
	MMC	37.94	27.98	34.61	19.62	42.73	14.89	20.75
	Monguno	40.83	36.83	38.67	14.50	35.83	13.08	16.17
Yobe	Bade	43.67	20.00	20.33	23.67	47.00	16.33	11.00
	Damaturu	39.96	26.77	28.19	21.27	48.69	15.81	14.85
	Potiskum	39.83	29.25	29.00	21.83	41.92	15.50	19.58
Average	All	41.29	25.48	27.96	20.26	47.61	15.25	16.64

A. Self-Efficacy

The distribution of the scores in the self-efficacy scale is skewed to the left with ceiling effects, reflecting that teachers in the sample tend to feel capable of performing their duties as teachers regarding engaging parents and students and having a classroom climate conducive to learning. Scores for self-efficacy group towards the higher end of the scale, with over 80% of the scores in the range from 40 points to the maximum scale score, 50. This, along with the data anomalies found in the community of Biu, suggest that there might be a tendency to provide socially desirable responses. It does not seem to be an intrinsic problem of the scale, as evidenced by the low scores from the communities of Konduga and Jere. Overall scale scores have a negative skew (leaning to the right), indicating a self-report of high self-efficacy scores in the surveyed teachers. The mean for the scale scores is at over 80% of the maximum scale score, indicating a high prevalence of very high scores for this scale.

B. Job Satisfaction

The distribution of the Job Satisfaction scale is considerably skewed to the left and with ceiling effects, indicating a prevalence of high scores. Particularly in this scale, the most prevalent score is the top score for the scale (24 points), which might be indicative of a ceiling effect. It is to be noted

that this is a very short scale that relates only to general job satisfaction, and not to specific issues related to satisfaction with teaching as a job. As it is, the scale manages to make visible a small cluster of teachers reports very low levels of satisfaction as seen in the first bar form left to right in the histogram for this scale: these teachers are located in the community of Monguno, which includes 3 schools and 9 teachers.

C. Moods and Feelings

The scale has 18 questions, each using a scale of 5 for a theoretical maximum score of 90 points, corresponding to the highest levels of depression-related behaviors. Average scores fall towards the low end of the scale, with a mean score of 30 points, corresponding to 30% of the maximum scale score and a median of only 26 points. The maximum score recorded was 76 points, and very few teachers report such figure. In fact, the vast majority of teachers report very low scores on this scale, which has a right-skewed distribution, which seems indicative of low levels of depression-related behaviors for the sample. This includes both physical and emotional manifestations of distress, which were used as prompts. The highest levels for this scale seem to be focused on three communities in the state of Borno: Konduga, MMC and Monguno.

D. Burnout

In this scale, where higher scores indicate higher burnout, such a low mean score might be indicative of a floor effect. Teachers in the community of Monguno show the highest scores for burnout in the sample. They seem to be an anomaly in a context of low burnout scores. Either there is a focalized problem in this regard at Monguno, or perhaps the scale might need some adjustment.

E. Associated scale: Coaching Experiences

The scale has 12 items with a maximum scale score codified as 5. Hence the maximum scales score is 60 points. As seen in the histogram of scale scores, the majority of teachers self-reported very high scores for their coaching experience. Even more, teachers not participating in the IRC tutoring program, reported slightly higher scores (average = 52) than the participating teachers (average = 49). The difference, however, is not statistically significant ($t = 0.29$, 179 d.f., $p > 0.05$). As in other scales, the mean is very close to 80% of the maximum score indicating the high prevalence of high scores in the results.

F. Associated scale: Social Support and Collaboration

The scale has 5 items with scale of four points for a maximum score of 20 points. As seen in the score distributions and similarly to other scales, scores are grouped towards the higher end of the scale. This scale inquiries about the frequency with which teachers meet with other teachers to plan instructional activities, observe colleagues, or work together with other teachers. Teachers in the sample report that they do this daily or almost daily.

G. Associated scale: Teaching Barriers

The scale has 6 items with scale of three points and items refer to teachers' perceptions on how much different factors (lack of prerequisite knowledge, disruptive students, and other teaching

barriers) prevented them from teaching their classes. The scale showed no clear skewing, with a relatively uniform distribution of scores across the spectrum of the scale, which granted a maximum score of 24 points.

5.2 Is there evidence of reliability?

Reliability for the scales related to Teacher Wellbeing was high, even after the deletion of anomalous response patterns corresponding to some schools and communities. ***Table 5*** below presents summary results for each of the subscales included in the study.

Table 5 Reliability and inter-item correlations for the scales in the study

Measure	Number of Items	Cronbach's alpha	Average interitem correlation
1. Self-efficacy	10	0.93	0.59
2. Job satisfaction	4	0.93	0.77
3. Moods and feelings	18	0.94	0.45
4. Burnout	20	0.93	0.40
5. Coaching experience	12	0.93	0.54
6. Social Support and Collaboration	5	0.91	0.66
7. Teaching Barriers	6	0.57	0.18

All of the scales show substantial reliability values for Cronbach's alpha (Shrout & Lane, 2012), with the exception of the Teaching Barriers scale, which is well below the conventional threshold of 0.7. Recommended values for interitem correlations are between 0.15 to 0.50 (Clark & Watson, 1995). In this regard, some of the scales show higher than recommended values. Notably, the average interitem correlation for the scale of Job Satisfaction is 0.77, suggesting that there might be higher than desirable redundancy in the items in the scale; this is of particular concern because the scale is only 4 items-long. In this regard, lower values for inter-item correlations for the scale of Teacher Self-Efficacy would be desirable, as well as for other scales not directly part of TWB such as Social Supports and Collaboration. Detailed reliability statistics for each scale are available in Annex 8.1, including item-test correlations and the effect of removing each item in Cronbach's alpha.

5.3 Is there evidence of construct validity?

Following Messick's interpretation of validity (Maul, 2018), finding evidence for construct validity involves gathering different sources of evidence supporting that the scales jointly measure aspects integral to a particular construct, in this case TWB, and behave following theory's expected patterns relative to each other and relative to constructs that are external to TWB. Implied in this discussion is the fact that each individual construct behaves as theory indicates, i.e.; that it has a certain structure as defined in the literature. This section presents the evidence for the individual scales that

are proposed to be different aspects of TWB: self-efficacy, job satisfaction, burnout and moods and feelings. In the following section (5.4) it will be discussed how constructs behave relative to each other. The Discussion section addresses construct underrepresentation and construct irrelevant variance.

5.3.1 Self-Efficacy

The hypothesis that items in this scale conformed to a 2-factor solution for items related to teacher skills (3, 4, 5, 7, and 8) and for items related to class management (1, 2, 6, 9,10) was tested by running a CFA model by the method of maximum likelihood. In this case, the model did not converge, meaning that the hypothesized structure does not find support in the data. Hence, a simpler model where all indicators depend on a single latent variable (Teaching Efficacy) was tested. This model converged and produced high loadings (all close to 0.8) and low residuals, as seen in the Table 6 below. Fit indices, however, are below acceptable thresholds (Brown, 2015), although not extremely and might improve with a different and/or larger sample. More studies with this scale are required.

Table 6. Self-Efficacy loadings and residuals

Item Text	Loading	Residual
I have the ability to get parents involved in their child's education	0.747	0.44
I can make my classroom a safe place for students, both emotionally and physically	0.868	0.25
I feel confident about my abilities as a teacher	0.871	0.24
If a student does not remember information, I gave in a previous lesson, I would know how to help them remember	0.792	0.37
When a student gets a better grade than he or she usually gets, it is because I found a better way	0.801	0.36
If a student in my class is undisciplined, I know some techniques to redirect him or her	0.803	0.36
I can get through to even the most difficult or unmotivated students	0.726	0.47
I can motivate students who show low interest in school	0.825	0.32
I can get students to work in groups or pairs	0.815	0.34
I can help students overcome some difficult home and community conditions	0.810	0.34

Fit indices: CFI = 0.915, TLI = 0.891, RMSEA = 0.143 ($p < 0.05 = 0$), and SRMR = 0.042.

5.3.2 Job Satisfaction

The scale is very short scale (4 items) and displays very high inter-item correlations, as shown in Annex 8.1. It does not seem to add analytic value to consider the internal structure of such a short scale of very similar items. A single latent variable model for all the items shows high loadings, low residual, and good fit statistics (Table 7). All values for the fit statistics are conventionally good. (Brown, 2015)

Table 7. Job Satisfaction loadings and residuals

Item Text	Loading	Residual
I am happy to come to work	0.91	0.17
I want to continue for a long time in my current workplace	0.95	0.09
My current job is rewarding	0.78	0.39
I enjoy being in my current job position	0.86	0.26

Fit indices: CFI = 0.984, TLI = 0.953, RMSEA = 0.166 ($p < 0.05 = 0.013$), and SRMR = 0.022.

5.3.3 Moods and Feelings

The items in the scale can be classified as behavioral and physical manifestations of depression. This classification resulted in a CFA model where despite loadings being not so low, some of the residuals are high (Table 9), and more importantly, CFA fit indices are well below conventional thresholds.

Table 8. Job Stress Fit loadings and residuals

Item Text	Loading	Residual
Been lacking energy?	0.61	0.62
Felt that you have lost interest in your usual activities?	0.78	0.40
Felt that you have lost confidence in yourself?	0.66	0.56
Felt hopeless?	0.69	0.53
Felt that you have difficulty concentrating?	0.71	0.50
Felt that you have lost weight (due to poor appetite)?	0.79	0.38
Been waking up early?	0.59	0.65
Felt slowed up?	0.74	0.45
Tended to feel worse in the morning?	0.75	0.43
Felt very anxious?	0.65	0.57
Been worrying a lot?	0.80	0.36
Irritable?	0.67	0.54
Been having difficulty relaxing?	0.74	0.46
Been sleeping poorly?	0.75	0.44
Been having headaches or neck aches?	0.74	0.46
Been having any of the following: Trembling, Tingling, Dizzy spells, Sweating, Diarrhea, or Needing to pass water more often than usual?	0.65	0.57
Been worrying about your health?	0.50	0.55
Been having difficulty falling asleep?	0.67	0.56

Fit indices: CFI = 0.757, TLI = 0.722, RMSEA = 0.146 ($p < 0.05 = 0$), and SRMR = 0.082.

5.3.4 Burnout

The burnout scale has a relatively high number of items (20) and has been reported previously in the literature several years ago by Richmond and Gorham (1992). The theoretical grouping of items

shown in section 3.3.4 and solutions involving between 2 and 4 factors were used as input for a CFA model. However, all of the candidate models derived from theoretical considerations or EFA results showed both poor loadings and high residuals, as seen in Table 8 below, showing results for a model with two hypothesized factors: internal and external sources of burnout. Fit statistics are far from conventional thresholds of acceptability.

Table 9. Burnout loadings and residuals

Item Text	Loading	Residual
I am bored with my job	0.734	0.462
I am tired of my students	0.887	0.214
I am weary with all of my job responsibilities	0.671	0.550
My job doesn't excite me any more	0.753	0.433
I dislike going to my job	0.626	0.608
I feel alienated at work	0.735	0.460
I feel frustrated at work	0.603	0.637
I avoid communication with students	0.830	0.310
I avoid communication with my colleagues	0.694	0.518
I communicate in a hostile manner at work	0.589	0.653
I feel ill at work	0.581	0.662
I think about calling my students ugly names	0.914	0.165
I avoid looking at my students	0.511	0.739
My students make me sick	0.589	0.653
I feel sick to my stomach when I think about work	0.401	0.840
I wish people would leave me alone at work	0.540	0.708
I dread going into a classroom	0.680	0.537
I am apathetic about my job	0.715	0.488
I feel stressed at work	0.714	0.490
I have problems concentrating at work	0.539	0.710

Fit indices: CFI = 0.646, TLI = 0.602, RMSEA = 0.175 ($p < 0.05 = 0$), and SRMR = 0.097.

5.4 Is there evidence of convergent and discriminant validity?

Convergent validity refers to the idea that different measures of a common construct should exhibit "high levels of agreement" or converge to a common truth. For example, a positive aspect of TWB like job satisfaction should show a negative correlation with a negative aspect of the construct, like burnout. Discriminant validity refers to the fact that measures of different constructs should not be as highly correlated as to render redundant one construct relative to another. For instance, while a positive correlation is expected between self-efficacy and job satisfaction, the correlation should have a moderate magnitude. Table 10 shows how the correlations obtained between total scores for the scales compare with the hypotheses presented in Section 4.

Results are color-coded, showing differences between the predicted relationships and the expected results. Large differences are shown in orange, medium differences in yellow, and coincidences in green. A difference was considered large if both its magnitude and direction were opposite to the prediction. A difference was considered medium if there was a coincidence in the direction/sign of the correlation, but a difference in the size of the correlation that is not large, i.e., it involves adjacent categories in correlation size, with correlation values close to the boundary between categories (e.g., medium a strong are adjacent but weak and strong are not). Coincidence (or small difference) was determined when magnitude and direction were the same or when despite being different, the magnitude of the correlation was close to the threshold between adjacent correlation size categories. Note that correlation sizes will be determined following guidelines by Courtney (2018) and Mukaka (2012), and the thresholds used for low, moderate, and high Pearson correlation coefficients are |0.3|, |0.5|, and |0.7|.

Table 10 Obtained correlations between scale scores

Scale	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1. Self-Efficacy	1						
2. Job Satisfaction	H=Strong, positive O=Weak, negative (-0.07)	1					
3. Burnout	H=Strong, negative O=Weak, negative (-0.08)	H=Strong, negative O=Weak, negative (-0.12)	1				
4. Moods and Feelings	H=Moderate, negative O=Weak, negative (-0.16)	H=Negative O=Weak, negative (-0.07)	H=Strong, positive O= Moderate, positive (0.61)	1			
5. Coaching Experience	H=Weak, not negative O= Moderate, positive (0.64)	H=Weak O=Weak (0.06)	H=Moderate, negative O=Weak, negative (-0.24)	H=Weak, not positive O= Weak, not positive (-0.3)	1		
6. Social Support and Collaboration	H=Weak, not negative O=Moderate, positive (0.39)	H=Weak, positive O=Weak, positive (0.13)	H=Moderate, negative O=Weak (0.1)	H=Weak, not positive O=Weak, not positive (-0.03)	H=Weak O=Weak (0.25)	1	
7. Teaching Barriers	H=Moderate, negative O=Weak, negative	H=Moderate, negative	H=Moderate, positive O=Weak, positive (0.14)	H=Moderate, positive	H=Weak O=Weak (0.1)	H=Weak, positive O=Weak,	1

	(-0.10)	O=Weak, positive (0.22)		O=Moderate, positive (0.34)	positive (0.1)

Note: H = Hypothesis or expected correlation, O = Obtained correlation in analysis

In general terms, the predicted correlations between the constructs hypothesized to be part of TWB differ importantly from what was predicted. Out of 6 correlations, 2 show a large difference with their prediction, and 4 show a medium difference. Self-efficacy is the construct that departs the most from what was expected in terms of correlations, with no predictions matching the results closely. For example, the correlation with Job Satisfaction was expected to be strong and positive, while a weak and negative value was obtained. The correlations with the Burnout and Moods and Feelings scales were expected to be strong and negative, while the data showed a negative, but weak correlation close to zero. For the Job Satisfaction scale, the correlation obtained with Burnout was in the same direction as expected (negative), but while the expected correlation was strong, the obtained correlation was weak (-0.12). The scale of Moods and Feelings correlated with the other components of TWB closer to the hypothesis proposed than other constructs. The scale shown two close matches with Self-efficacy and Job Satisfaction (for which a negative relationship was expected, and a negative and close to zero correlation was obtained) and a coincidence with Burnout (expected to be positive and strong with a 0.61 value obtained).

With regards to convergent validity for TWB, all of the correlations match in direction but differ in magnitude, except for one (the correlation between self-efficacy and job satisfaction). We believe that the proper estimation of the magnitude of these correlations would require lager samples. Hence, this initial evidence is a least, in the right direction.

Discriminant validity refers to the fact that measures of different constructs should not be as highly correlated as to render redundant one construct relative to another. For instance, while a positive correlation is expected between self-efficacy and job satisfaction, the correlation should have a moderate magnitude. The same empirical findings that weaken the hypothesis of convergent validity in our results provide good evidence for discriminant validity since none of the constructs involved in TWB correlates as strongly to each other so as to render it redundant or unnecessary. The largest correlation between two constructs is between moods and feelings and burnout, with a value of 0.61, which although high, is far from being at the level (over 0.8) that would indicate that one of the measures is redundant.

5.5 Is there evidence of concurrent validity?

In this report, we establish concurrent validity by comparing to TWB constructs that, while not integral to TWB, such as those included in the Barriers to Teaching, Social Support and Collaboration and Coaching Experience, are attributes related to it (West & Beckman, 2018) and analyzing the correlations between the scales (Table 10, rows 5, 6 and 7)

The Coaching Experience scale correlations Job Satisfaction and Moods and Feelings were as precited, while the relationship with Burnout was not as strong as expected, though in the predicted

direction. The greatest departure from the predicted relationship was with Self-efficacy, which was predicted as weak and non-negative, while the data showed that it was positive and in the high-end of moderate (0.64). The Social Support and Collaboration scale showed a better pattern with two coincidences (with Job Satisfaction, weak and positive; and Mood and Feelings, weak, non-negative) and two partial matches (with Self-Efficacy and Burnout) with a match regarding direction (positive and negative, respectively), but not with the expected magnitudes. The Teaching Barriers scale's predicted correlations showed a good match with the Moods and Feelings scale (with a moderate, positive correlation, as predicted), and coincidence with Self-Efficacy and Burnout in terms of direction, though differing in magnitude. Finally, the moderate negative correlation expected between Teaching Barriers and Job Satisfaction was contradicted by findings, which show a weak, positive relationship (0.2).

6. Discussion and Recommendations

Generating evidence around TWB – its relationship with relevant student outcomes, what works to effectively support it, how, for whom, and at what cost – can help inform both programs and sector-level policies to support quality education and student learning. However, high-quality evidence starts with scales that have evidence of reliability and validity, and that have been previously used and validated in specific contexts. In the present study, we analyzed the psychometric properties of four scales measuring TWB that, while based on previously existing scales, lacked evidence of reliability and validity in contexts of conflict and crisis, such as northeast Nigeria.

This study thus contributes to the literature by testing and providing psychometric evidence of previously used scales, building evidence of validity and reliability for these, and adding to the limited availability of measures used in crises and conflict-affected contexts. While the data collected in the context of an education program allowed for this study, the implemented scales also posed certain limitations. First, two scales, Job Satisfaction and Moods and Feelings, were not explicitly framed to inquiry about teaching and were rather general. Second, the data collected possessed certain anomalies that were discovered during the data analyses.

The results of the study are promising, though the psychometric evidence of the scales is mixed. At face value, and according to definitions explored in the literature, the scales seem to measure the constructs they are intended to measure. All the scales show good reliability, evidence on convergent and discriminant validity is in the right direction, and the scales do not correlate too strongly with each other as to render them redundant. However, construct validity analyses resulted in a lack of fit (or below the acceptable thresholds for CFI and TLI of at least 0.95) for all the models specified except for the Job Satisfaction scale.

Given the mixed results, in this section, we discuss plausible explanations and limitations of the study and detail recommendations for refining the scales and future analyses. First, implementation of the scales should be strengthened to ensure data integrity, as there were indications of data anomalies in our study sample. Further, the response patterns of certain scales also suggested

potential ceiling effects, see Self-Efficacy and Job Satisfaction, or floor effects, see Burnout scale, which could negatively impact CFA fit statistics. We suggest addressing these issues, first, by analyzing inter-rater reliability during data collection and second, by encouraging teachers to be candid in their responses. Data collection teams must insist and ensure clarifying the lack of repercussions (negative or positive) of teachers' responses that might seem problematic or different from what they think evaluators want to hear. Finally, testing the scales with a new sample could improve fit statistics for CFA models of each individual construct and provide a more solid ground for testing construct validity.

On the concurrent validity, while we do not observe a complete coincidence between predictions and results between the TWB and the associated scales, contradictions were not prevalent, and two of the three supporting scales also show adequate reliability values. Because scales do not correlate too strongly with each other as to render them redundant, we believe construct underrepresentation might be more salient than construct irrelevance variance, and the inclusion of general indicators of wellbeing (such as Job Satisfaction, and Moods and Feelings) does not represent construct irrelevant variance. We believe this strengthens the case for the TWB tool being valid but requiring further refinement and better-quality data.

Given the positive reliability results, we suggest testing the scales with larger samples which could provide stronger evidence for convergent and discriminant validity, as correlations as highly dependent on sample sizes. The limitations of the evidence on construct validity also indicate that the scales' items wording could be further revised to focus more explicitly on teachers' experiences. Further, as more research on the relationship between these constructs stems from conflict and crisis contexts, the hypotheses about the correlations among constructs could be revisited.

For example, based on our initial literature review, we hypothesized the correlation between the Self-Efficacy and Job Satisfaction scale would be strong and positive, but results resulted in a small negative correlation of -0.07. While one plausible explanation could be the small sample size, another is the lack of specificity of the Job Satisfaction scale with the teaching profession. A more nuanced reframing of the items and a larger sample could result in a result aligned with our initial hypothesis. In addition, research about how self-efficacy behaves in contexts of conflict and crises, could help revise the initial hypothesis and help guide a more concerted and systematic exercise of contextualization of the tools. Importantly, subsequent contextualization should also pay special attention to linguistic and contextual relevance and include the voices of teachers and relevant stakeholders. While the version of the scales used were based on scales previously used by TWB research, to our knowledge, these were not pilot tested, with the Nigerian population of teachers. We strongly suggest investing in a deeper exercise of contextualization for the particular contexts to be used and review the scales' items to produce stronger and more appropriate scales. Finally, an additional consideration should be made to the linguistic appropriateness of the scales. While all the teachers in the study's sample were fluent in English, implementing tools in respondents' Mother Tongue is key on the contextualization process.

The lack of specificity and contextualization of scales to the context and purpose, might entail that certain relevant aspects in the conceptualization of teacher wellbeing are left unexplored, hence effectively configuring a possible stance of under-representation. Our study, thus supports Hascher & Waber, (2021) recommendation of avoiding the use of general concepts of teacher wellbeing and rather addressing the specific challenges and demands of the teaching profession. Our punctual recommendation to improve the current measures, without the threat of increasing construct-irrelevant variance, is that TWB scales include both general and more specific items related not only to the teaching profession but to the context of education in emergencies.

7. References

APA. (2020). *Self-Efficacy Teaching Tip Sheet*. Https://Www.Apa.Org. https://www.apa.org/pi/aids/resources/education/self-efficacy

Bakker, A. B., & Demerouti, E. (2007). The Job Demands-Resources Model: State of the Art. Journal of Managerial Psychology, 22, 309-328.
https://doi.org/10.1108/02683940710733115

Bandura, A. (2006). Guide for Constructing Self-Efficacy Scales. In *Self-Efficacy Beliefs of Adolescents*.

Bermejo-Toro, L., Prieto-Ursúa, M., & Hernández, V. (2016). Towards a model of teacher well-being: Personal and job resources involved in teacher burnout and engagement. *Educational Psychology*, 36(3), 481-501. https://doi.org/10.1080/01443410.2015.1005006

De Biagi, N. B., Celeri, E. H., & Renshaw, T. L. (2017). Technical adequacy of the teacher subjective wellbeing questionnaire with Brazilian educators. *Journal of Psychoeducational Assessment*, 36(8), 850–855.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (Second edition..). The Guilford Press.

Bujang, M. A., Omar, E. D., & Baharum, N. A. (2018). A Review on Sample Size Determination for Cronbach's Alpha Test: A Simple Guide for Researchers. *The Malaysian Journal of Medical Sciences : MJMS*, *25*(6), 85–99. https://doi.org/10.21315/mjms2018.25.6.9

Buric, I., Sliskovic, A., & Penezic, Z. (2019). Understanding teacher well-being: A cross-lagged analysis of burnout, negative student-related emotions, psychopathological symptoms, and resilience. *Educational Psychology*, 39(9), 1136–1155.

Burns, R. A., & Machin, M. A. (2013). Employee and workplace well-being: A multi-level analysis

Capone, V., & Petrillo, G. (2018). Mental health in teachers: Relationships with job satisfaction, efficacy beliefs, burnout and depression. *Current Psychology*, 39(5), 1757–1766.

Clark, L. A., & Watson, D. (1995). *Constructing Validity: Basic Issues in Objective Scale Development*. 11.

Courtney, M. R. C. (2018). Pearson Correlation Coefficient. In *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. SAGE Publications, Inc. https://doi.org/10.4135/9781506326139

de Winter, J. C. F., Dodou, D., & Wieringa, P. A. (2009). Exploratory Factor Analysis With Small Sample Sizes. *Multivariate Behavioral Research*, *44*(2), 147–181. https://doi.org/10.1080/00273170902794206

DeVellis, R. F. (2012). *Scale development: Theory and applications* (3rd ed..). SAGE.

Diazgranados, S & Lee, J (2019) Choosing and Contextualizing Assessment Measures in Educational Contexts, International Rescue Committee. https://inee.org/measurement-library/measure-guidance-choosing-and-contextualizing-assessment-measures-educational-contexts

Fabrigar, L., & Kan, M. P. H. (2018). Exploratory Factor Analysis. In *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. SAGE Publications, Inc. https://doi.org/10.4135/9781506326139

Falk, D., Varni, E., Finder, J., Frisoli, P., & Frisoli, J. (2019). *Landscape Review: Teacher Well-being in Low Resource, Crisis, and Conflict-affected Settings*. Education Equity Research Initiative. https://inee.org/sites/default/files/resources/TWB%20Landscape%20Review_August%202019_0.pdf

Gkolia, A., Belias, D., & Koustelios, A. (2014). *TEACHER'S JOB SATISFACTION AND SELF-EFFICACY: A REVIEW*. 23.

Goldberg D, Bridges K, Duncan-Jones P, Grayson D. Detecting anxiety and depression in general medical settings. British Medical Journal 1988;297:897-899

Granziera, H., Collie, R., Martin, A. (2021). Understanding Teacher Wellbeing Through Job Demands-Resources Theory. In: Mansfield, C.F. (eds) Cultivating Teacher Resilience. Springer, Singapore. https://doi.org/10.1007/978-981-15-5963-1_14

Hair, J. F. (Ed.). (2010). *Multivariate data analysis* (7th ed). Prentice Hall.

Hall-Kenyon, K. M., Bullough, R. V., MacKay, K. L., & Marshall, E. E. (2014). Preschool teacher well-being: A review of the literature. *Early Childhood Education Journal*, 42(3), 153–162.

Hascher, T., & Waber, J. (2021). Teacher wellbeing: A systematic review of the research literature from the year 2000–2019. *Educational Research Review*, *34*, 100411. https://doi.org/10.1016/j.edurev.2021.100411

Ho, C.-L., & Au, W.-T. (2006). Teaching Satisfaction Scale: Measuring Job Satisfaction of Teachers. *Educational and Psychological Measurement*, *66*(1), 172–185. https://doi.org/10.1177/0013164405278573

Hobson, A. J., & Maxwell, B. (2017). Supporting and inhibiting the wellbeing of early career secondary school teachers: Extending self-determination theory. *British Educational Research Journal*, *43*(1), 168–191. https://doi.org/10.1002/berj.3261

INEE. (2022a). *INEE - A Guidance Note for Teacher Wellbeing in Emergency Settings*. Inter-agency Network for Education in Emergencies (INEE) c/o International Rescue Committee.

INEE. (2022b). *Measurement Library | INEE*. https://inee.org/measurement-library

INEE. (2022c). *We-Act | INEE*. https://inee.org/measurement-library/we-act?webform_id=toolkit_resources

Jennings, P. A. (2016). CARE for teachers: A mindfulness-based approach to promoting teachers' social and emotional competence and well-being. In K.A. Schonert-Reichl & R.W. Roeser (Eds.). *Handbook of mindfulness in education* (pp. 133-148). New York, NY: Springer.

Jennings, P. A., Brown, J. L., Frank, J. L., Doyle, S., Oh, Y., Davis, R., ... & Greenberg, M. T. (2017). Impacts of the CARE for Teachers program on teachers' social and emotional competence and classroom interactions. *Journal of Educational Psychology*, 109(7), 1010.

Jordans, M., Coetzee, A., & Steen, F. (2019). *We-Act Measure: Feasibility and Sensitivity to Change Testing among Service Providers in Gaza*. 17.

Kim, A. R. and H. (2019, May 6). To promote success in schools, focus on teacher wellbeing. *Brookings*. https://www.brookings.edu/blog/education-plus-development/2019/05/06/to-promote-success-in-schools-focus-on-teacher-well-being/

Kirk, J., & Winthrop, R. (2013). Teaching in contexts of emergency and state fragility. In J. Kirk, M. Dembélé, & S. Baxter (Eds.) *More and better teachers for quality education for all: Identity and motivation, systems and support* (pp. 121-139).

Knapp, T. (2011). *Validity*. Sage Publications, Inc. https://doi.org/10.4135/9781412963947

Laine, S., Tossavainen, K., Pertel, T., Lepp, K.. Isoaho, H., Saaranen, T., (2018). Occupational Well-Being: A Structural Equation Model of Finnish and Estonian School. *Global Journal of Health Science,* 10 (8) , 79-90. 10.5539/gjhs.v10n8p79.

McCallum, F., & Price, D. (2010). Well teachers, well students. *The Journal of Student Wellbeing*, 4(1), 19-34.

Maul, A. (2018). Validity. In *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. SAGE Publications, Inc. https://doi.org/10.4135/9781506326139

Mendenhall, M. (2017). Strengthening teacher professional development: Local and global communities of practice in Kakuma refugee camp. London, Save the Children, UNHCR and Pearson.

Mukaka, M. (2012). A guide to appropriate use of Correlation coefficient in medical research. *Malawi Medical Journal : The Journal of Medical Association of Malawi*, *24*(3), 69–71.

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum Sample Size Recommendations for Conducting Factor Analyses. *International Journal of Testing*, *5*(2), 159–168. https://doi.org/10.1207/s15327574ijt0502_4

Naghieh, A., Montgomery, P., Bonell, C. P., Thompson, M., & Aber, J. L. (2015). Organisational interventions for improving wellbeing and reducing work-related stress in teachers. *Cochrane Database of Systematic Reviews*, *4*. https://doi.org/10.1002/14651858.CD010306.pub2

OECD (2020) Teachers' well-being: A framework for data collection and analysis. OECD Education Working Paper No. 213

Randall, J., & Hung, H. J. (2018). Confirmatory Factor Analysis. In *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. SAGE Publications, Inc. https://doi.org/10.4135/9781506326139

Renshaw, T. L., Long, A. C., & Cook, C. R. (2015). Assessing teachers' positive psychological functioning at work: Development and validation of the Teacher Subjective Wellbeing Questionnaire. *School Psychology Quarterly*, 30(2), 289–306.

Richmond, V. P., & Gorham, J. (1992). *Communication, learning, and affect in instruction*. Burgess International Group.

Shrout, P. E., & Lane, S. P. (2012). Reliability. In H. Cooper (Ed.), *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics.* (pp. 643–660, Chapter xliv, 744 Pages). American Psychological Association (Washington, DC, US). http://dx.doi.org.proxy.bc.edu/10.1037/13619-034

Skaalvik, E. M., & Skaalvik, S. (2007). *Dimensions of Teacher Self-Efficacy and Relations With Strain Factors, Perceived Collective Teacher Efficacy, and Teacher Burnout*. 15.

Skaalvik, E. M., & Skaalvik, S. (2011). Teachers' feeling of belonging, exhaustion, and job satisfaction: The role of school goal structure and value consonance. *Anxiety, Stress & Coping*, *24*(4), 369–385. https://doi.org/10.1080/10615806.2010.544300

Skaalvik, E. M., & Skaalvik, S. (2014). Teacher Self-Efficacy and Perceived Autonomy: Relations with Teacher Engagement, Job Satisfaction, and Emotional Exhaustion. *Psychological Reports*, *114*(1), 68–77. https://doi.org/10.2466/14.02.PR0.114k14w0

Soares, F., Menezes Cunha, N., & Frisoli, P. How Do We Know If Teachers Are Well? The Wellbeing Holistic Assessment for Teachers Tool. *Journal on Education in Emergencies* (2021)

Tavakol, M. (2018). Coefficient Alpha. In *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. SAGE Publications, Inc. https://doi.org/10.4135/9781506326139

Ter Meulen W.G., Draisma S., van Hemert A. M., et al. (2021) Depressive and anxiety disorders in concert-A synthesis of findings on comorbidity in the NESDA study. *Journal of Affective Disorders*.

Thakur, M., Chandrasekaran, V., & Guddattu, V. (2018). Role conflict and psychological well-being in school teachers: A cross-sectional study from Southern India. *Journal of Clinical and Diagnostic Research*, 12(7), 1–6.

Thapa, A., Cohen, J., Guffey, S., & Higgins-D'Alessandro, A.

Tubbs-Dolan, C & Weiss-Yagoda, J (2019). Education in Emergencies Evidence for Action: Measurement and Metrics Initiative. 3EA, International Rescue Committee and NYU Global TIES for Children

Veronese, G., Pepe, A., Dagdukee, J., & Yaghi, S. (2018). Teaching in conflict settings: Dimensions of subjective wellbeing in Arab teachers living in Israel and Palestine. *International Journal of Educational Development*, 61, 16–26.

West, C., & Beckman, T. (2018). Concurrent Validity. In *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. https://doi.org/10.4135/9781506326139

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety. *Educational and Psychological Measurement*, *76*(6), 913–934. https://doi.org/10.1177/0013164413495237

Wolf, S., Torrente, C., McCoy, M., Rasheed, D., & Aber, J. L. (2015). Cumulative risk and teacher well-being in the Democratic Republic of the Congo. Comparative Education Review, 59(4), 717-742.

8. Annexes

8.1 Scale reliability statistics by item

A. Teacher Self-Efficacy

Prompt: Choose one option that best describes you
Scale reliability: 0.93
Average inter-item correlation: 0.59

Item Identifier	Item Text	Obs	Item-test	Item-rest	Average inter-item correlation	Cronbach's alpha if item is removed
TEFF_1	I have the ability to get parents involved in their child's education	151	0.68	0.59	0.61	0.93
TEFF_2	I can make my classroom a safe place for students, both emotionally and physically	151	0.85	0.80	0.58	0.92
TEFF_3	I feel confident about my abilities as a teacher	153	0.85	0.80	0.58	0.92
TEFF_4	If a student does not remember information, I gave in a previous lesson, I would know how to help them remember	151	0.80	0.74	0.59	0.93
TEFF_5	When a student gets a better grade than he or she usually gets, it is because I found a better way	150	0.77	0.71	0.59	0.93
TEFF_6	If a student in my class is undisciplined, I know some techniques to redirect him or her	150	0.81	0.76	0.58	0.93
TEFF_7	I can get through to even the most difficult or unmotivated students	149	0.74	0.67	0.60	0.93
TEFF_8	I can motivate students who show low interest in school	146	0.80	0.74	0.59	0.93
TEFF_9	I can get students to work in groups or pairs	144	0.80	0.75	0.58	0.93
TEFF_10	I can help students overcome some difficult home and community conditions	142	0.81	0.76	0.58	0.93

B. Job Satisfaction

Prompt: Choose one option that best describes you.
Sale reliability: 0.93
Average inter-item correlation: 0.77

Table 11 Reliability Statistics for the Job Satisfaction Scale

Item Identifier	Item Text	Obs	Item-test correlation	Item-rest correlation	Average inter-item correlation	Cronbach's alpha if item is removed
JOBSTF_1	I am happy to come to work	179	0.92	0.86	0.76	0.90

JOBSTF_2	I want to continue for a long time in my current workplace	179	0.94	0.89	0.74	0.89
JOBSTF_3	My current job is rewarding	179	0.87	0.77	0.82	0.93
JOBSTF_4	I enjoy being in my current job position	179	0.91	0.84	0.77	0.91

C. Moods and Feelings

Prompt: During the past month, how often have you...?
Scale reliability: 0.94
Average inter-item correlation: 0.45

Item Identifier	Item Text	Obs	Item-test correlation	Item-rest correlation	Average inter-item correlation	Cronbach's alpha if item is removed
MOODFL_1	Been lacking energy?	166	0.68	0.62	0.45	0.93
MOODFL_2	Felt that you have lost interest in your usual activities?	161	0.77	0.73	0.44	0.93
MOODFL_3	Felt that you have lost confidence in yourself?	164	0.65	0.60	0.45	0.93
MOODFL_4	Felt hopeless?	164	0.68	0.63	0.45	0.93
MOODFL_5	Felt that you have difficulty concentrating?	161	0.71	0.67	0.45	0.93
MOODFL_6	Felt that you have lost weight (due to poor appetite)?	164	0.77	0.73	0.44	0.93
MOODFL_7	Been waking up early?	163	0.55	0.48	0.46	0.94
MOODFL_8	Felt slowed up?	162	0.70	0.66	0.45	0.93
MOODFL_9	Tended to feel worse in the morning?	162	0.76	0.72	0.44	0.93
MOODFL_10	Felt very anxious?	163	0.64	0.59	0.45	0.93
MOODFL_11	Been worrying a lot?	161	0.78	0.74	0.44	0.93
MOODFL_12	Irritable?	161	0.71	0.66	0.45	0.93
MOODFL_13	Been having difficulty relaxing?	164	0.71	0.67	0.45	0.93
MOODFL_14	Been sleeping poorly?	163	0.72	0.68	0.45	0.93
MOODFL_15	Been having headaches or neck aches?	161	0.74	0.70	0.45	0.93
MOODFL_16	Been having any of the following: Trembling, Tingling, Dizzy spells, Sweating, Diarrhea, or Needing to pass water more often than usual?	161	0.65	0.60	0.45	0.93
MOODFL_17	Been worrying about your health?	149	0.55	0.49	0.46	0.94
MOODFL_18	Been having difficulty falling asleep?	148	0.71	0.67	0.45	0.93

Note: Item MOODFL_7 (Been waking up early?) does not seem to have a clear directionality regarding the construct (not clear if high stress results in waking up early); hence the item is candidate for removal from the scale.

D. Teacher Burnout

Prompt: Choose one option that best describes you
Scale reliability: 0.93

Average inter-item correlation: 0.40

Item Identifier	Item Text	Obs	Item-test correlation	Item-rest correlation	Average inter-item correlation	Cronbach's alpha if item is removed
BRNOUT_1	I am bored with my job	147	0.68	0.64	0.40	0.93
BRNOUT_2	I am tired of my students	145	0.75	0.71	0.40	0.93
BRNOUT_3	I am weary with all of my job responsibilities	139	0.68	0.61	0.40	0.93
BRNOUT_4	My job doesn't excite me any more	146	0.67	0.63	0.40	0.93
BRNOUT_5	I dislike going to my job	147	0.67	0.62	0.40	0.93
BRNOUT_6	I feel alienated at work	144	0.67	0.62	0.40	0.93
BRNOUT_7	I feel frustrated at work	144	0.65	0.61	0.41	0.93
BRNOUT_8	I avoid communication with students	146	0.75	0.71	0.40	0.93
BRNOUT_9	I avoid communication with my colleagues	146	0.68	0.64	0.40	0.93
BRNOUT_10	I communicate in a hostile manner at work	145	0.62	0.57	0.41	0.93
BRNOUT_11	I feel ill at work	146	0.59	0.54	0.41	0.93
BRNOUT_12	I think about calling my students ugly names	143	0.79	0.75	0.39	0.93
BRNOUT_13	I avoid looking at my students	147	0.54	0.49	0.41	0.93
BRNOUT_14	My students make me sick	145	0.62	0.57	0.41	0.93
BRNOUT_15	I feel sick to my stomach when I think about work	145	0.57	0.52	0.41	0.93
BRNOUT_16	I wish people would leave me alone at work	131	0.54	0.48	0.41	0.93
BRNOUT_17	I dread going into a classroom	131	0.79	0.76	0.39	0.93
BRNOUT_18	I am apathetic about my job	130	0.72	0.68	0.40	0.93
BRNOUT_19	I feel stressed at work	131	0.66	0.61	0.40	0.93
BRNOUT_20	I have problems concentrating at work	131	0.62	0.57	0.41	0.93

E. Coaching Experience

Prompt: To what extent, does your coach...
Overall scale reliability: 0.94
Average inter-item correlation: 0.54

Item Identifier	Item Text	Obs	Item-test correlation	Item-rest correlation	Average inter-item correlation	Cronbach's alpha if item is removed
COACHX_1	let you know what she/he expects from you?	150	0.64	0.57	0.56	0.93
COACHX_2	coach give you constructive criticism?	149	0.65	0.58	0.56	0.93
COACHX_3	let you know why you did not do something well?	147	0.72	0.66	0.55	0.93
COACHX_4	talk to you about your behavior?	149	0.69	0.62	0.56	0.93
COACHX_5	stand by you?	147	0.84	0.80	0.53	0.93

COACHX_6	cheer you up?	149	0.81	0.76	0.54	0.93
COACHX_7	guide you in a right direction?	148	0.82	0.78	0.53	0.93
COACHX_8	give you good advice?	147	0.82	0.78	0.53	0.93
COACHX_9	tell you to persevere?	150	0.81	0.77	0.54	0.93
COACHX_10	comfort you?	147	0.78	0.73	0.54	0.93
COACHX_11	help you to clarify your problems?	150	0.83	0.79	0.53	0.93
COACHX_12	reassure you?	150	0.75	0.69	0.55	0.93

F. Social Supports and Collaboration

Prompt: How often do you...?
Scale reliability: 0.90
Average inter-item correlation: 0.65

Item Identifier	Item Text	Obs	Item-test correlation	Item-rest correlation	Average inter-item correlation	Cronbach's alpha if item is removed
SSCOL_1	discuss how to teach a particular topic with other teachers?	173	0.82	0.72	0.67	0.89
SSCOL_2	collaborate in planning and preparing instructional materials with other teachers?	174	0.89	0.82	0.62	0.87
SSCOL_3	share what you have learned about your teaching experiences with other teachers?	172	0.87	0.79	0.64	0.87
SSCOL_4	visit another classroom to learn more about teaching with other teachers?	171	0.79	0.67	0.69	0.90
SSCOL_5	work together to try out new ideas with other teachers?	174	0.87	0.79	0.64	0.88

G. Teaching Barriers

Prompt: Swipe to answer to what extent the following limit how you taught your class
Scale reliability: 0.57
Average inter-item correlation: 0.18

Item Identifier	Item Text	Obs	Item-test correlation	Item-rest correlation	Average inter-item correlation	Cronbach's alpha if item is removed

TEACHBAR_1	Students lacking prerequisite knowledge of skill	179	0.54	0.28	0.19	0.54
TEACHBAR_2	Students suffering from lack of basic nutrition	179	0.54	0.28	0.19	0.54
TEACHBAR_3	Students suffering from not enough sleep	179	0.63	0.40	0.16	0.49
TEACHBAR_4	Students with special needs (e.g Physical disabilities, mental or emotional/psychological impairment)	179	0.64	0.41	0.16	0.48
TEACHBAR_5	Disruptive Students	179	0.35	0.06	0.25	0.63
TEACHBAR_6	Uninterested students	179	0.68	0.46	0.14	0.46