



ASSESSING CLASSROOM QUALITY IN CRISIS CONTEXTS: PSYCHOMETRIC REPORT OF THE TEACHER CLASSROOM OBSERVATION (TCO) INSTRUMENT.

Technical Working Paper

May 25, 2024

This report was prepared by Autumn Brown, former Technical Advisor at the IRC and current PhD student at the University of Pennsylvania and Dr. Jeongmin Lee, former Researcher for Education at the International Rescue Committee (IRC), currently working at the University of Notre Dame. We thank the teachers and enumerators in Tanzania who participated in the data collection. We also thank Porticus, the Medical Research Council, and an anonymous donor who made this study possible.

Suggested Citation:

Brown, A. & Lee, J. (2024). *Assessing Classroom Quality in Crisis Contexts: Psychometric Report of the Teacher Classroom Observation (TCO) Instrument [Working Paper]*. International Network for Education in Emergencies.

Table of Contents

INTRODUCTION	4
TCO OVERVIEW	5
Theoretical background	5
Development procedure	5
TCO items	6
TANZANIA CONTEXT	7
METHODS	8
Sample	8
Procedure	8
Analysis	10
Item characteristics	10
Interrater agreement	10
Construct validity	11
Composite reliability	12
RESULTS	12
Item characteristics	12
Interrater agreement	15
Construct validity	16
Composite reliability	17
RECOMMENDATIONS	19
Recommendations for content adaptation	19
Recommendations for expanding the evidence base of TCO's psychometric qualities	19
Recommendations for assessor training	20
Recommendations Scoring the TCO	21
CONCLUSION	22
REFERENCES	23

ASSESSING CLASSROOM QUALITY IN CRISIS CONTEXTS: PSYCHOMETRIC REPORT OF THE TEACHER CLASSROOM OBSERVATION (TCO) INSTRUMENT

ABSTRACT

The availability of psychometrically robust instruments to evaluate teacher practice and classroom quality in low- and middle-income countries is limited, particularly in contexts of conflict and crisis. The International Rescue Committee (IRC) and NYU Global TIES for Children iteratively developed the Teacher Classroom Observation (TCO) tool to address this gap. This report investigates the psychometric properties of the TCO tool in a sample of 150 teachers in the Nyarugusu refugee camp in western Tanzania. Specifically, we show (1) item characteristics, (2) interrater reliability, (3) construct validity, and (4) composite reliability.

We find empirical evidence to support the TCO's validity and reliability in a refugee context, highlighting its utility in generating actionable evidence. In this report, we detail the development process of the TCO and provide recommendations for researchers and practitioners interested in utilizing this tool to improve education interventions in similar contexts.

INTRODUCTION

Research worldwide suggests that the quality of teachers' practice is the most influential factor in student learning, from pre-kindergarten to secondary education, and a teacher's impact has effects into adulthood (Azam & Kingdon, 2015; Burchinal et al., 2008; Chetty et al., 2014; Hanushek & Rivkin, 2010; Howie, 2005). Snilstveit and colleagues (2016) reviewed 216 programs in 52 developing countries and reported that among varying types of programs to improve learning (e.g., school grants, technology-assisted instruction, community monitoring) those that supported teachers in developing effective instructional strategies had the greatest and most consistent positive outcomes. Despite the importance of high-quality teaching, teaching practice is rarely monitored in low- and middle-income countries (Ladics et al., 2018), and structural factors—teacher characteristics; e.g., number of years of experience and education level—are used as proxies for teacher quality, despite evidence that these structural factors only weakly associate with teaching quality (Burchinal et al., 2002). In contexts of conflict and crisis, challenges in assessing teacher practice, including a lack of valid and reliable measures that can be administered quickly, force practitioners to rely on structural indicators, limiting practitioners' and researchers' ability to better understand how to support teacher practice and improve programs. This report examines if Teacher Classroom Observation (TCO) could be a useful tool that fills this gap.

As we detail below, the TCO was designed to assesses the quality of teacher practice and the classroom environment for monitoring and evaluation purposes. The TCO was co-developed by the Internal Rescue Committee (IRC) and the Global TIES for Children Center at New York University (NYU) and its development has been an iterative and collaborative process spanning six countries over the course of five years. The design of the TCO was specific to conflict and crisis settings, both in the constructs the instrument is designed to capture and in the administration of the enumerator training. Different iterations of TCO have been used in Africa, including Nigeria, Sierra Leone, Niger, and Tanzania, and the Middle East, including Lebanon and Iraq. The data collected from each country have been utilized for an incremental and iterative improvement of TCO, including refinement of item content, wording, scoring metrics and rubrics, and training materials. In this report, we specifically show the TCO's psychometric properties using data from Tanzania, where the most recent pilot occurred, focusing on (1) item characteristics, (2) interrater agreement, (3) construct validity, and (4) composite reliability.

TCO OVERVIEW

Theoretical background

The International Rescue Committee developed the Learning in a Healing Classroom (LIHC) curricula to meet the needs of students in contexts of conflict and crisis. The LIHC approach combined the available evidence on how teachers can support children exposed to adversity with 30 years of experience implementing education programs in emergency contexts (International Rescue Committee, 2018). LIHC focuses on improving children's reading and math skills while supporting their social and emotional development through safe, predictable, and emotionally supportive classrooms. The items of the TCO were developed to reflect the LIHC teacher training and the specific practices taught to teachers. Specifically, the general teaching practices taught to teachers, classroom management techniques designed to create a safe and predictable classroom, and techniques to create an emotionally supportive environment—dimensions that are reflected in the TCO constructs (see [IRC Creating Healing Classrooms Facilitator Guide](#)). The goal of the TCO development was to create a valid and reliable measure of teacher practice that could be used across IRC programming alongside LIHC. Additionally, we aimed to provide a *feasible* instrument that could be implemented despite the intense time and financial constraints typical of emergency programming.

Development procedure

The first iteration of the TCO was developed in Lebanon by the Lebanon Education Technical Team. The instrument was designed as a monitoring and evaluation tool to use for donor reporting. In 2015, the IRC and NYU Global TIES for Children began implementing a research program funded by Dubai Cares (see [Education in Emergencies: Evidence for Action](#)) in Lebanon, Niger, and Sierra Leone to assess the impact of LIHC. Understanding the quality of teacher practice was part of this effort and the TCO was designed to reflect the LIHC training in all three countries. After the initial testing in these countries, a new iteration was tested in Iraq. Given the ultimate goal of being able to use the instrument globally, the adaptations were to ensure cross-country applicability and reliability. Each enumerator training included on-site inter-rater reliability testing to improve the training materials and clarity of the items. The

Tanzania version of the TCO reviewed in this report is the fourth, and most recent, version of the instrument and training materials.

TCO items

The TCO contains 22 items across four discrete dimensions of teacher instruction and classroom management quality. They include (a) Time-on-Task, (b) Teaching Practices (TEACH), (c) Classroom Management (MANAGE), and (d) Emotional Supportiveness (SUPPORT). The Time-on-task dimension differs in that it is a measure of the amount of instructional time spent teaching and learning course concepts, i.e., time spent on academic tasks, as opposed to non-academic tasks such as greetings, unscheduled interruptions, or material distribution. Enumerators note start and end times for non-academic activities, add those durations, and subtract them from the total instructional time.

The Teaching Practices (TEACH) construct has seven items. It measures aspects of teacher pedagogical practices. The Classroom Management (MANAGE) construct contains five items. It focuses on the routines the teacher employs to structure the classroom activities, create a predictable environment, and ensure children feel safe while staying on task. Lastly, the Emotional Supportiveness (SUPPORT) construct consists of five items. It measures the extent of emotional support teachers provide and an emotionally safe environment.

Table 1. TCO Items

TEACH	T1	Feedback
	T2	Checking for Understanding
	T3	Supplemental Materials
	T4	Opportunities to Practice
	T5	Questioning Techniques
	T6	Connecting to Students Lives
	T7	Engaging Different Types of Learners
MANAGE	M1	Lesson Objective
	M2	Classroom Rules
	M3	Discipline Consistency
	M4	Use of Classroom Space
	M5	Positive Discipline Strategies

SUPPORT	E1	Calm Tone of Voice
	E2	Positive Words
	E3	Promoting Self Compassion
	E4	Recognizing Perseverance
	E5	Encouraging Positive Behavior

All items are on a 4-point Likert scale where “1 = No evidence of the behavior/negative behavior, 2 = Attempted, but poorly executed, 3 = Good effort, 4 = Exemplary.” To ensure consistency and transparency in rating, each item has a description of the behavior in general as well as descriptions of what the practice would look like as a 1, 2, 3, or 4.

Field administration

The original TCO was used for donor reporting purposes and the inter-rater reliability was not tested. In order to make the instrument available for research purposes, the IRC and NYU Global TIES developed materials for training raters. One consistent piece of feedback received throughout the development process was that observing classrooms was difficult, time-consuming, confusing, and overly technical. To understand if this could be improved solely through the design of the materials, the IRC worked with user-centered designer Sarah Fathallah in Lebanon to observe the TCO training, collect feedback from participants, and suggest improvements. The results were used to inform both the design of the instrument and training materials to increase user-friendliness.

The improved training was shorter (2.5 days compared to 4 days) and included videos for raters to watch and score so that inter-rater reliability could be assessed in real time and challenging items could be clarified before the data collection. The training also included interactive and energizing activities to allow participants to quickly familiarize themselves with each item and learn about cognitive bias and the necessity of maintaining impartiality while observing teachers. (See *TCO Facilitator’s Guide*)

TANZANIA CONTEXT

Nyarugusu refugee camp is located in Kigoma, Tanzania, and was established in 1996 for refugees fleeing the ongoing conflict in the DRC and Burundi. The latest UNHCR data shows that the camp hosts approximately 150,000 refugees of whom over half (55%) are children under

the age of 17. The Tanzanian government restricts refugee access to services outside of the camp, necessitating NGO provision of health, education, and food services.

The IRC is the NGO responsible for education and child protection in Nyarugusu and supports 27 schools with the majority of programming focused on teacher professional development and training. One aspect of the teachers' training is emotional self-regulation and positive discipline practices to reduce violence against children in the form of corporal punishment in schools through a program called Preventing Violence Against Children in Schools (PVACs). PVACs is a collaboration between the IRC and research partners at the London School of Health and Tropical Medicine (LSHTM), Innovations for Poverty Action (IPA), the National Institute for Medical Research (NIMR), and the Behavioral Insights Team (BIT) to assess how the intervention might shift teachers' beliefs, attitudes, and behaviors related to corporal punishment. The intervention consisted of 10-week peer-led sessions where teachers learned strategies and alternatives to corporal punishment, including positive discipline and classroom management techniques. Given the lack of validated teacher-classroom observation tools in similar contexts, PVAC decided to use, and adapt, the TCO to assess teacher practice.

METHODS

Sample

150 teachers randomly selected from the treatment and control groups were observed using the TCO. Observed classrooms spanned all grades and subjects, but over half of the observations came from grades 1-6. Teachers were also split between Congolese and Burundian nationalities.

Procedure

Raters were trained over 2.5 days (see *Table 2 for the training agenda*). The first day of the training included reviewing each item of the tool and rating a series of hypothetical scenarios followed by general training on conducting observations, including an overview of cognitive bias and how these biases can limit the impartiality necessary to rate properly.

Table 2: TCO 4.0 Training Agenda

Tanzania TCO Training Agenda			
Day 1	8:30 - 9:00	30min	Session 1: Welcome and Introduction

	9:00 - 9:15	15min	Session 2: Why Teacher Classroom Observation?
	9:15 - 9:45	45min	Session 3: Getting to Know the Tool
	9:45 - 10:00	15min	Break
	10:00 - 11:00	60min	Session 4: What Makes a Good Observer?
	11:00 - 12:15	75min	Session 5: Becoming Objective
	12:15 - 13:15	60min	Lunch break
	13:15 - 13:30	15min	Session 6: Using the Tool for Measurement
	13:30 - 14:45	75min	Session 7: Scoring Guide Review
	14:45 - 15:00	15min	Break
	15:00 - 15:45	45min	Session 7: Scoring Guide Review (continued)
	15:45 - 16:15	30min	Conclusion
Day 2	8:30 - 9:30	60min	Session 8: Review Recap
	9:30 - 10:00	30min	Session 9: Observation Protocol
	10:00 - 10:15	15min	Break
	10:15 - 12:00	135min	Session 10: Practice
	12:00 - 13:00	60min	Lunch break (and inter-rater reliability testing)
	13:00 - 14:30	90min	Session 10: Practice (continued)
	14:30 - 14:45	15min	Break (and inter-rater reliability testing)
	14:45 - 15:30	45min	Session 10: Practice (continued)
	15:30 - 16:00	30min	Conclusion
Day 3	9:00 - 9:30	30min	Session 11: Welcome and Introduction
	9:30-10:30	1hr	Session 12: Practice
	10:30-10:45	15 mins	Break
	10:45-12:30	1hr45	Session 12: Practice and Conclusion

Ten enumerators were hired to participate in the TCO training to become raters. All ten enumerators were refugees from Nyarugusu and had experience as data collectors, some as recently as the PVAC baseline data collection. The enumerators all spoke English fluently and were a mix of Burundian and Congolese. The training was co-facilitated by the IRC technical advisor leading the TCO development and the Education Lead in Nyarugusu, Tanzania. Both trainers spoke English and French fluently, allowing for easier communication on especially confusing items.

The day before the training, IRC staff received consent from teachers to film their classrooms and videos were filmed and selected for training purposes based on the quality of the

teaching. Videos were uploaded onto tablets and tablets were administered to the trainees. After each rater watched a video on their tablet, they would rate using the TCO instrument and submit their ratings to the trainer. The trainer would input the ratings into an Excel sheet (See *Interrater Reliability Testing Sheet*) to see which items had the least amount of agreement. They would then spend time going over the item in detail with the trainees before moving on to another video. In the end, six raters “passed” based on their consistent level of agreement over the practice videos and these eight raters were selected to observe teachers. Raters were randomly paired and then geographically distributed with one pair covering a certain number of schools. The raters would sit separately in the back of the classroom and individually complete their observation and upload their TCO scores without consulting each other.

Analysis

Item characteristics

To examine item characteristics of the TCO, we computed various types of descriptive statistics such as the means, standard deviations, minimum, and maximum in order to evaluate the distribution of the items and assess if any TCO items showed a floor or ceiling effect. Floor or ceiling effects can indicate that these items are not sensitive to variation in teacher practice.

Interrater agreement

We assessed the interrater agreement of TCO by using two agreement measures: Krippendorff’s alpha coefficient and the intra-class correlation (ICC) coefficient. Krippendorff’s alpha is one of the most generalized measures of agreement. It quantifies the degree of agreement by comparing observed disagreement with the expected level of disagreement by chance. It is applicable to data on various levels of metrics (e.g., nominal, ordinal, interval, and more), any number of observers, and incomplete or missing data (Hayes & Krippendorff, 2007; Krippendorff, 2011).

Krippendorff’s alpha scores range from zero to one, with zero indicating a lack of agreement beyond chance (i.e., agreement is the same as would be expected due to chance) and one indicating perfect agreement (Krippendorff, 2011). In this paper, we interpreted Krippendorff’s alpha values, following the thresholds of agreement suggested by Krippendorff (2018) in Table 3.

The ICC measures interrater agreement based on the ratio of true score to observed score variance. There are various versions of the ICC, and the version applicable to the present study is a one-way random, absolute agreement model (Bandalos, 2018). As such, two (but not always the same two) enumerators were assigned to each LF (one-way). These enumerators were reviewed to be representative of a larger pool of enumerators so that we could generalize our study's results to other enumerators (random). ICC values typically range from zero to one, with zero indicating the absence of agreement and one indicating perfect agreement (LeBreton & Senter, 2008). We interpreted ICC values, following the thresholds of agreement suggested by Portney and Watkins (2015) in Table 3.

Table 3. Cutoffs applied to interpret interrater agreement statistics

Level of agreement	ICC (Portney & Watkins, 2015)	Krippendorff's alpha (Krippendorff, 2018)
Needs improvement	< 0.5	< 0.67
Fair	0.50 – 0.75	n/a
Good	0.75 – 0.90	0.67 - 0.80
Excellent	0.90 – 1.00	0.80 - 1.00

Construct validity

To test the factorial validity of TCO, we tested how well our data fit the hypothesized structure of TCO, focusing specifically on the following three constructs: TEACH (7 items), MANAGE (5 items), and SUPPORT (5 items). In determining parameter estimation methods, we recognized that all TCO's items were ordered-categorical items. We thus employed robust weighted least squares with mean and variance adjustment (WLSMV) as it is robust to the non-normality of categorical distributions (Xia & Yang, 2018). We evaluated model-data fit based on several fit statistics along with the thresholds of goodness of fit. The fit statistics included root mean square error of approximation (RMSEA), comparative fit index (CFI), Tucker-Lewis index (TLI), and standardized root mean square residual (SRMR), along with χ^2 statistics with $p > 0.05$. Table 4 displays the fit statistics we used, along with the conventional thresholds of goodness of fit.

Table 4. Goodness of fit statistics applied

Fit indices	Good fit	Acceptable fit
--------------------	-----------------	-----------------------

Comparative Fit Index (CFI)	> 0.95	> 0.92
Tucker-Lewis Index (TLI)	> 0.95	> 0.92
Root mean square error of approximation (RMSEA)	< 0.06	< 0.08
Standardized root mean square residual (SRMR)	< 0.05	< 0.08

Source: Finney & DiStefano, 2013; Hu & Bentler, 1999

Composite reliability

When construct validity was established, we assessed the composite reliability of each of the factors contained in TCO—a measure of the aggregate reliability of scale items (Huck, 2011)—by using McDonald’s omega (ω). In this report, omega values > 0.70 were interpreted as a good level of composite reliability. Values between 0.60-0.70 were indicative of an acceptable level (Lance et al., 2006). Additionally, we also use Cronbach’s alpha internal consistency value for TCO users who seek to use the sum or average scores for formative purposes.

RESULTS

Item characteristics

Overall, the TCO data were well spread out across the four-point scale for most items except for a few items. These items are Time-on-Task, TEACH item 3, and MANAGE items 2 and 3. Time-on-Task is negatively skewed, whereas the other items are positively skewed. These patterns, by and large, indicate a likely ceiling effect for Time-on-task and a floor effect for the rest of the items. In other words, the Time-on-task item is not able to distinguish fine-grained differences in teachers’ abilities even though such differences may exist. However, given that these items are more a measure of *dosage* than quality, we do not see this as inconsistent with the instrument’s intended use.

On the other hand, the data are normally distributed at the factor level for all three factors measuring quality—TEACH, MANAGE, and SUPPORT. Table 5 presents the mean, standard deviation, and minimum and maximum scores of TCO 4.0 items and factors. Figures 1 and 2 visualize score distributions at the item and factor levels.

Table 5. Mean, standard deviation, and distribution of LFs by item and factor

Variables	Mean	SD	Minimum	Maximum
-----------	------	----	---------	---------

<i>Item level</i>					
Time on task (%)		0.96	0.05	0.74	1
TEACH	T1	2.1	0.9	1	4
	T2	2.4	0.9	1	4
	T3	1.2	0.7	1	4
	T4	2.3	0.9	1	4
	T5	2.3	0.8	1	4
	T6	1.8	1.0	1	4
	T7	1.9	0.7	1	4
MANAGE	M1	1.5	0.7	1	4
	M2	1.1	0.5	1	4
	M3	1.5	0.9	1	4
	M4	3.0	0.9	1	4
	M5	2.0	1.0	1	4
SUPPORT	E1	3.0	0.9	1	4
	E2	2.3	0.9	1	4
	E3	1.8	0.7	1	3
	E4	1.8	0.9	1	4
	E5	2.6	0.9	1	4
<i>Factor level</i>					
TEACH		1.9	0.5	1	3.3
MANAGE		1.9	0.5	1	4
SUPPORT		2.3	0.6	1	3.7

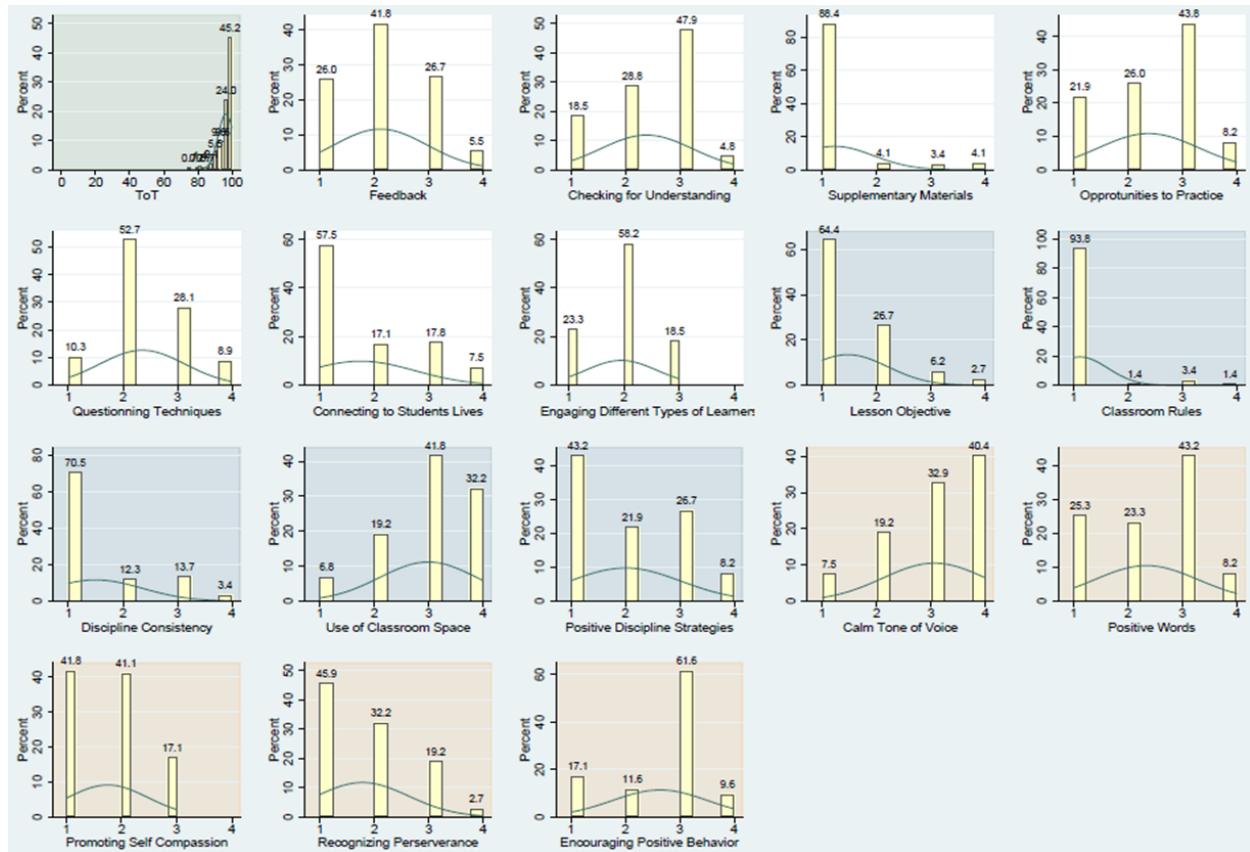


Figure 1. Score distribution by item

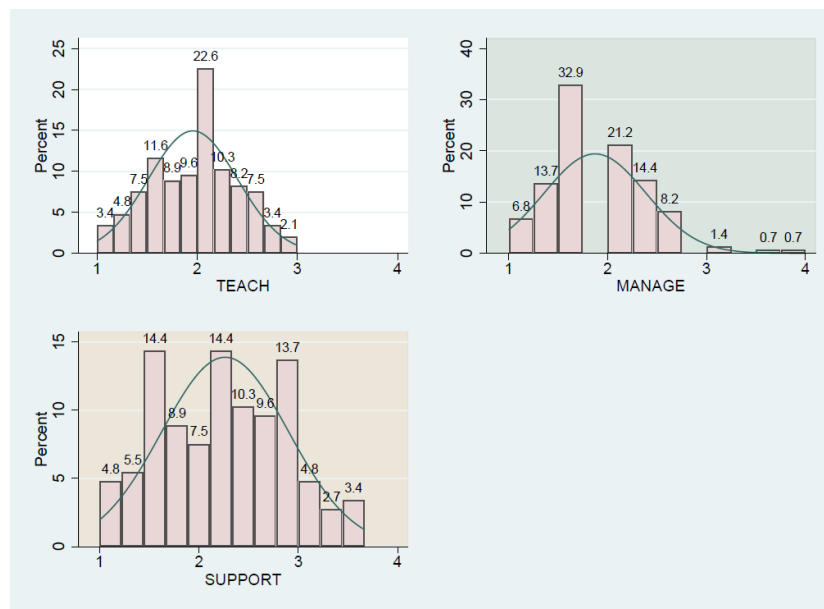


Figure 2. Score distribution by factor

Interrater agreement

All TCO items and factors showed a good to excellent level of interrater agreement. At the item level, the alpha values ranged from 0.79 (T1) to 0.95 (T2 and M5). When scores were aggregated at the factor level, all factors established an excellent level of agreement, with alpha scores ranging from 0.85 for MANAGE to 0.94 for SUPPORT. The result was consistent when the TCO data were assessed with the ICC. Table 6 shows the Krippendorff's alpha and ICC values of TCO items and factors.

Table 6. Krippendorff's alpha and intra-class correlation values of TCO items and factors

Variables		Krippendorff's Alpha	Agreement tenable? (Yes = ✓: No = ✗)	ICC	Agreement tenable? (Yes = ✓: No = ✗)
<i>Item level</i>					
Time on task (%)		0.88	✓	0.94	✓
TEACH	T1	0.79	✓	0.79	✓
	T2	0.95	✓	0.96	✓
	T3	0.86	✓	0.86	✓
	T4	0.81	✓	0.80	✓
	T5	0.88	✓	0.90	✓
	T6	0.81	✓	0.81	✓
	T7	0.80	✓	0.89	✓
MANAGE	M1	0.85	✓	0.83	✓
	M2	0.87	✓	0.89	✓
	M3	0.80	✓	0.83	✓
	M4	0.90	✓	0.89	✓
	M5	0.95	✓	0.95	✓
SUPPORT	E1	0.83	✓	0.86	✓
	E2	0.90	✓	0.89	✓
	E3	0.80	✓	0.80	✓
	E4	0.84	✓	0.83	✓
	E5	0.74	✓	0.81	✓
<i>Factor level</i>					
TEACH		0.91	✓	0.96	✓
MANAGE		0.85	✓	0.94	✓
SUPPORT		0.94	✓	0.97	✓

Construct validity

In Table 7, we provide an overview of the statistics we used to evaluate how well the three-factor measurement model we hypothesized for TCO fit the data. The first model with all 17 items (T1-7, M1-5, and E1-5) showed a poor model-data fit as all fit statistics were below our acceptance thresholds, as shown in the second column of Table 12. We investigated inter-item correlations and loading coefficients alongside modification indices to iterate our initial model. We noticed three items such as T3 “Supplemental Materials”, M2 “Classroom Rules”, and E1 “Calm Tone of Voice” were negatively and weakly correlated with other items within their respective factor. They also showed factor loadings lower than 0.5, which indicated a poor item (T3 = -0.23, M2 = 0.37, and E1 = 0.41) (Comrey & Lee, 2013; Fabrigar et al., 1999). These items were also the most heavily skewed. We thus removed the three items and ran a revised three-factor model.

The revised model still showed a poor model-data fit, as evidenced in the third column of Table 7. We investigated item qualities using the same approach used with the initial measurement model. We noticed two items such as T6 “Connecting to Students Lives” and M4 “Use of Classroom Space”, were poorly loaded on their corresponding factor ($\lambda = .26$ for T6 and $\lambda = 0.39$ for M4). We ran a reduced model with these two items removed.

This measurement model showed a good model-data fit, as evidenced in the last column of Table 7. Overall, all items were positively and strongly loaded onto their respective factor, with factor loadings ranging from 0.59 to 0.91. All factors were also positively and strongly correlated with one another, with factor correlations ranging from 0.55 to 0.80. In Figure 3, we provide a visual of this model.

Table 7. Model iterations and goodness of fit statistics

Fit index	3 Factor model with all items	3 factor model w/o T3, M2, E1	3 factor model w/o T3, T6, M2, M4, E1
χ^2	238.33* (df = 116)	162.03* (df = 74)	76.42* (df = 51)
RMSEA	0.09 [0.07-0.10]	0.09 [0.07-0.11]	0.06 [0.03, 0.09]
CFI	0.90	0.93	0.98
TLI	0.88	0.91	0.97
SRMR	0.10	0.08	0.06
Overall fit	Poor	Poor	Good

Composite reliability

As the TCO established construct validity, we moved on to assess the composite reliability of its three factors. Overall, all these factors showed a good to an excellent level of reliability with $\omega = 0.85$ for TEACH (5 items including T1, T2, T4, T5, and T7), $\omega = 0.70$ for MANAGE (3 items including M1, M3, and M5), and $\omega = 0.87$ for SUPPORT (4 items including S2, S3, S4, and S5). These statistics indicate homogeneity among TCO items within each of its three factors. These items are consistent with one another and measure the same construct. Cronbach's alpha values show overall consistent results, with 0.80 for TEACH, 0.65 for MANAGE, and 0.78 for SUPPORT.

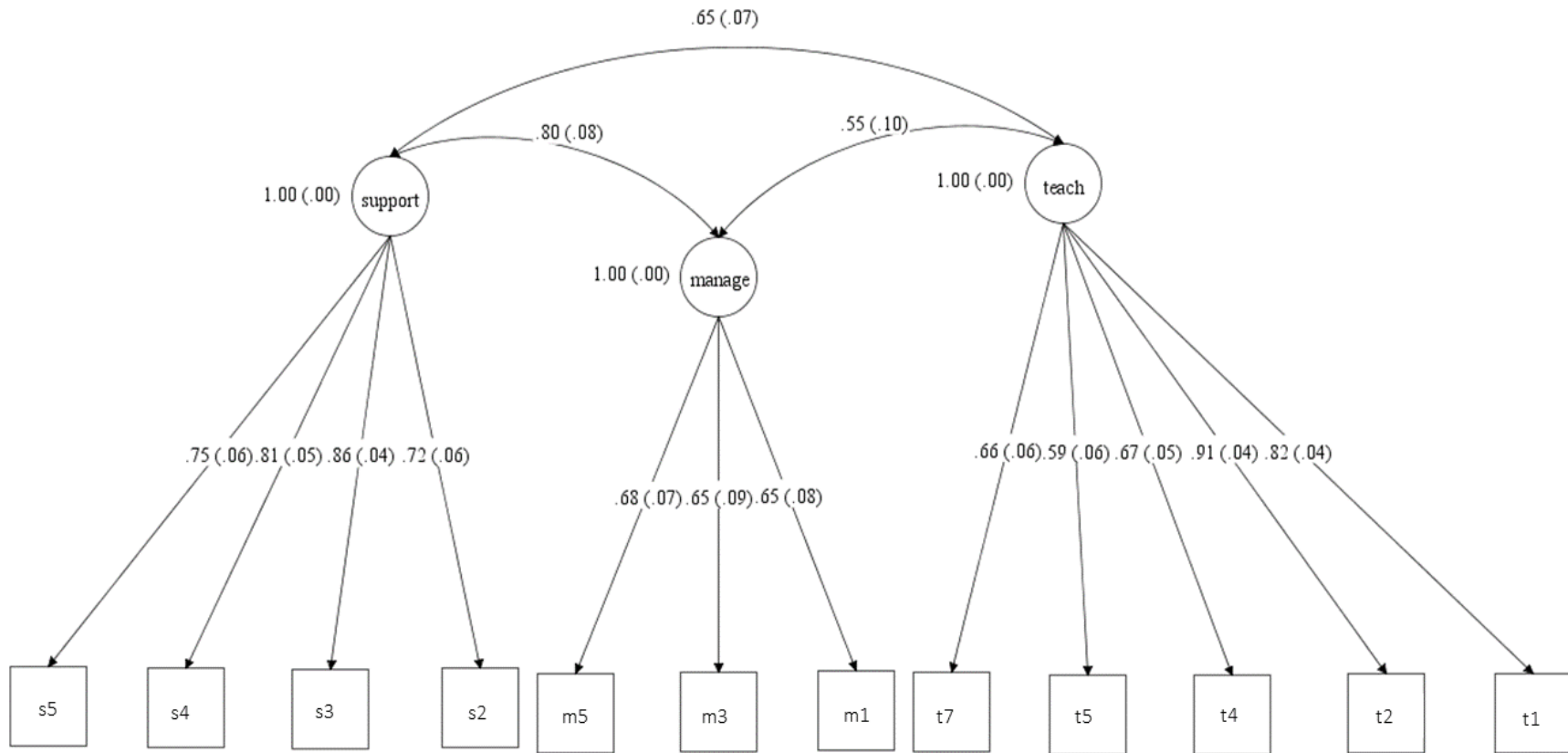


Figure 3. Standardized loadings and factor correlations of a three-factor TCO model

RECOMMENDATIONS

TCO demonstrates potential as a reliable and valid instrument for measuring the quality of teaching practices in low-resource and fragile contexts. To use TCO in other locations, both within and outside Tanzania, we recommend interested users adapt the TCO to better align with their specific context. This adaptation process could involve refining the tool itself and enhancing assessor training.

Recommendations for content adaptation

- Our findings within the PVAC project context in Tanzania revealed that Items T3, T6, M2, M4, and S1 had weak factor loadings and/or negative item correlations. T3 “Supplemental Material” and M2 “Classroom Rules” may be more appropriate as part of a structural measure given these correspond with the physical presence of classroom items and less with teaching practice. Items T6 “Connecting to Students Lives”, M4 “Use of Classroom Space” and E1 “Calm Tone of Voice” should be refined in future iterations.
- When adapting TCO for use in different linguistic, cultural, or contextual settings, we recommend conducting thorough validation processes to ensure its validity and reliability in the new context. This may involve translating the instrument into the local language(s), culturally validating the content and wording of items to ensure relevance and appropriateness, and conducting sensitivity checks to identify any potential biases or misinterpretations—as back-translating and establishing face validity.
- Finally, engaging with local stakeholders, including teachers, students, and community members, could provide valuable insights and perspectives to inform the adaptation process and enhance TCO’s acceptability in the target context.

Recommendations for expanding the evidence base of TCO’s psychometric qualities

- The TCO can also be used as a monitoring and coaching tool, but any use that requires comparing across raters should be avoided unless the interrater reliability has been assessed. Relevant recommendations include ensuring inter-rater reliability and preventing coding drift. Specifically, we suggest interested users implement paired

- coding sessions, where two or more coders independently evaluate the same classroom observations, and then compare their ratings to ensure consistency.
- Coaches in IRC’s programs have used the TCO, but more research could be conducted to understand how TCO’s psychometric properties fair over repeated use. We recommend that if using the TCO for coaching purposes, that data be kept with the coaches and comparisons across teachers should be avoided until further psychometric testing.
 - Further research could also extend the current report to test other aspects of TCO’s psychometric qualities, such as predictive and concurrent. Future studies could also test how the psychometric properties of this tool perform along multiple time points to strengthen the case for use as a monitoring tool.

Recommendations for assessor training

Our iterative testing and refinement of the TCO highlighted that the items and the instrument are only part of what makes for a high-quality assessment of teacher practice. We emphasize that enumerator training is critical to producing high-quality data and should be conducted with care. Hence, we recommend that interested users of the TCO incorporate a process for evaluating and enhancing the training materials to ensure the production of the highest-quality data possible.

- We highly recommend using videos taken from classrooms in the context where the data will be collected to be used to practice and test inter-rater reliability in real-time. Videos should be collected for the entire duration of a class period, filmed from a tripod in the back of the class so the entire classroom is within view, and have a strong enough microphone to pick up on everything being said in the classroom. Practice videos need to be a mix of high-quality, mixed, and low-quality practices.
- We suggest training a larger pool of enumerators and then selecting a subset that “passes” the training by scoring videos consistently with a master rater or peers.
- The repeated watching and scoring of the videos can become tedious, and it is recommended to break these sessions up over time if possible.
- The trained raters must collect the data immediately following the training while the process is still fresh in their minds. If time permits, a practice observation where

several raters go to the same classroom and observe and compare ratings is recommended.

- Finally, we want to highlight that regular refresher training sessions could also be conducted to reinforce coding criteria and standards, address any discrepancies, and update coders on any modifications to the instrument or evaluation process.

If users only want to use the TCO for formative assessment and coaching purposes, then they can use an abbreviated version of the TCO training materials (see TCO Facilitator's Guide).

However, if users decide to use the abbreviated training then we recommend they do not collect and use the data beyond individual coaching sessions as inter-rater reliability needs to be established in order to compare TCO scores across teachers and raters.

Recommendations Scoring the TCO

For researchers and evaluation: For researchers or those using the TCO for evaluation purposes, particularly in a new context, it is recommended to first establish factorial validity and reliability of the TCO instrument. If the measure is psychometrically sound, then we recommend that evaluators use factor scores to understand how teacher practice and classroom quality relate to other outcomes of interest.

For monitoring: Inter-rater reliability should be established before using the instrument to compare across teachers—even for monitoring purposes. However, for those using the instrument for monitoring or formative purposes, a mean score per section is sufficient. A mean score can be calculated by dividing the scores of each item by the total possible score.

For coaching: If coaches are using the TCO to guide their coaching practice, then scoring is not required at all or mean scores can be established and used at an individual teacher level (no comparisons across teachers).

CONCLUSION

Improving interventions to support children's positive development requires valid and reliable measures. Collecting observational data often demands highly specialized observers, extensive training, and the laborious task of coding hours of video footage. In emergency contexts with short timelines, practitioners and researchers need data quickly to learn and iterate. The TCO was developed specifically to address this need, providing a means to collect high-quality data on teacher practices in conflict and crisis settings. Our aim is that the TCO and its accompanying training materials will empower practitioners and researchers to access the information they need to design and implement interventions that support the needs of teachers and students. Moreover, we hope that future users will collectively expand the TCO's evidence base and best practices for implementation, ultimately producing high-quality data that effectively informs decision-making.

REFERENCES

- Azam, M., & Kingdon, G. G. (2015). Assessing teacher quality in India. *Journal of Development Economics*, 117, 74–83.
- Bingham, G. E., Quinn, M. F., & Gerde, H. K. (2017). Examining early childhood teachers' writing practices: Associations between pedagogical supports and children's writing skills. *Early Childhood Research Quarterly*, 39, 35–46.
- Burchinal, M., Howes, C., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Predicting child outcomes at the end of kindergarten from the quality of pre-kindergarten teacher–child interactions and instruction. *Applied Developmental Science*, 12(3), 140–153. <https://doi.org/10.1080/10888690802199418>
- Cameron, C. E., Connor, C. M., & Morrison, F. J. (2005). Effects of variation in teacher organization on classroom functioning. *Journal of School Psychology*, 43(1), 61–85.
- Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, 104(9), 2633–2679. <https://doi.org/10.1257/aer.104.9.2633>
- Cornelia, J., & Lainie, R. (2017). Conflict-sensitive teacher education: Viewing EDC's experience with the South Sudan teacher education project through a conflict-sensitive lens. *Journal on Education in Emergencies*, 1(1), 131–166.
- Dubai Cares. (2016). *Dubai Cares | Education in Emergencies: Evidence for Action (3EA) in Lebanon*. Dubai Cares. <https://www.dubaicare.ae/programs/education-emergencies-evidence-action-3ea-lebanon/>
- Finney, S. J., & DiStefano, C. (2013). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 439–492). Information Age Publishing, Inc.
- Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, 100(2), 267–271.
- Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1), 77–89.
- Howie, S. (2005). Contextual factors at the school and classroom level related to pupils' performance in mathematics in South Africa. *Educational Research and Evaluation*, 11(2), 123–140.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- INEE. (2016). *Teachers in crisis contexts training for primary school teachers*. Inter-agency Network for Education in Emergencies. <https://inee.org/resources/teachers-crisis-contexts-training-primary-school-teachers>
- International Rescue Committee. (2019). *IRC UK's Healing Classrooms | International Rescue Committee (IRC)*. <https://www.rescue.org/uk/irc-uks-healing-classrooms>

- Krippendorff, K. (2011). *Computing Krippendorff's alpha reliability*.
https://repository.upenn.edu/asc_papers/43
- Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). Sage publications.
- LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, 11(4), 815–852.
- Lee, J., & Zuilkowski, S. S. (2015). ‘Making do’: Teachers’ coping strategies for dealing with textbook shortages in urban Zambia. *Teaching and Teacher Education*, 48, 117–128.
<https://doi.org/10.1016/j.tate.2015.02.008>
- Lee, J., Zuilkowski, S. S., & D’sa, N. (2021). Organising primary grade literacy environments in Mozambique. *Learning Environments Research*, 24(2), 207–221.
<https://doi.org/10.1007/s10984-020-09327-w>
- Portney, L. G., & Watkins, M. P. (2015). *Foundations of clinical research: Applications to practice* (3rd ed.). Pearson.
- Richardson, E., MacEwen, L., & Naylor, R. (2018). *Teachers of refugees: A review of the literature*. Education Development Trust.
<https://www.educationdevelopmenttrust.com/EducationDevelopmentTrust/files/8e/8ebcf77f-4fff-4bba-9635-f40123598f22.pdf>
- Snilstveit, B., Stevenson, J., Menon, R., Phillips, D., Gallagher, E., Geleen, M., Jobse, H., Schmidt, T., & Jimenez, E. (2016). *The impact of education programmes on learning and school participation in low-and middle-income countries*. International Initiative for Impact Evaluation (3IE).
- Xia, Y., & Yang, Y. (2018). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-018-1055-2>