Scalability by Design: Leveraging DACUM and AI to Standardize Implementation Fidelity in Global Educational Interventions

Leslie Rosales

Universidad del Valle, Guatemala

Abstract

This research addresses the challenge of sustainable scalability in educational interventions by focusing on implementation fidelity as the cornerstone for success. The study argues that scientific precision in defining facilitator tasks is essential to replicate impact without compromising effectiveness. Utilizing the DACUM methodology, the author deconstructed staff responsibilities into measurable tasks to bridge the gap between theoretical models and classroom practice.
The study involves the validation of a monitoring instrument for Asociación Grupo Ceiba. Psychometric analyses, including Classical Item Analysis and Exploratory Factor Analysis, confirm the instrument is reliable and valid for assessing consistency. The findings also highlight Artificial Intelligence as a valuable tool for generating item banks, provided they undergo expert validation. Ultimately, this work provides a robust framework to ensure educational interventions remain effective as they scale.

Keywords: *Implementation fidelity, DACUM methodology, scalability, educational monitoring, psychometric validation, Artificial Intelligence.*

1. Introduction

The success of a high-quality educational intervention lies not only in its local results but also in its scalability; that is, its ability to replicate its impact on a large scale without compromising its original effectiveness [13]. However, this expansion process faces a critical challenge: ensuring that implementers faithfully execute the model while maintaining the flexibility needed to make appropriate adaptations to each context. In this sense, the sustainability of an intervention depends directly on the consistency and quality of its implementation over time and across different locations [4].
To achieve this balance, it is imperative to define with scientific precision what those responsible for implementing the program should do. This research argues that the foundation of successful implementation lies in a rigorous analysis of the responsibilities and tasks of the staff, specifically the teacher or facilitator, who acts as the final link in the delivery of educational services to the beneficiary [4].
Defining tasks is insufficient without a mechanism to verify the consistency of their implementation across different locations and over time. This technical need requires the design of a monitoring instrument that serves as an empirical bridge between the theoretically conceived educational model and teachers' practice. For this instrument to be a legitimate tool for decision-making, it must meet the criteria of validity and reliability [8].
Therefore, this research focuses on piloting and validating the monitoring instrument for the implementation of the flexible education model of Grupo Ceiba. This study used the **DACUM** (*Developing A Curriculum*) methodology as a starting point to break down the specific competencies and tasks of the facilitator within an educational model. Artificial Intelligence (AI) was also used for the comprehensive generation of observation items for the instrument.
Within this framework, the instrument's validity is based on the results derived from the DACUM process, ensuring that the items accurately represent the model's elements. Additionally, its reliability relies on translating these elements into observable tasks, allowing for an objective, consistent, and verifiable measurement of facilitators' performance in the field. This research demonstrates the validation of an instrument for monitoring the implementation fidelity of Grupo Ceiba's flexible education model through its group of facilitators.

1.1. Implementation Fidelity and Sustained Scalability

One of the fundamental premises for scaling an educational innovation is its proven effectiveness. However, when an innovation is not implemented as intended, the likelihood of it failing to achieve the expected results is high. This phenomenon underscores the critical need to measure the degree of

implementation fidelity before proceeding with any scaling efforts. An innovation may appear ineffective when the underlying problem is low fidelity, which raises doubts about the true causality of the results obtained [7].
Implementation fidelity is defined as the degree to which a practice is carried out according to the intended model, under the assumption that the success of any educational change depends on precise execution [7]. This concept is multidimensional and encompasses five key dimensions: dosage, adherence, responsiveness, quality, and differentiation. Of these, "quality" is particularly relevant to this research, as it includes the capacity for adaptive or flexible implementation depending on the context, evaluating the performance of the facilitator. Other dimensions analyze adherence to the model's components, the dosage of intervention received, user satisfaction, and the impact of the program's added value.
To measure these dimensions, especially in a complex educational model, subjective reports are insufficient. Century, Rudnick, and Freeman (2010) argue that direct observation is the most reliable and valid method for measuring fidelity, distinguishing between structural and instructional components [4]. However, for the observation to be rigorous, fundamental methodological questions arise: **what exactly will be observed in a facilitator while implementing the educational model? And are we confident that we are measuring implementation consistently and valid?**
It is at this point that the present research introduces the analysis of work tasks using the DACUM methodology. To answer the question of the degree of adaptive implementation of the educational model, it is imperative to break down the implementer's functions into observable tasks. The DACUM process provides the scientific basis for the validity of the measurement instrument, ensuring that what is observed at the sites coincides with the critical components of the original design. Furthermore, transforming these functions into concrete behavioral items ensures the reliability necessary to determine whether the variations at each site are valid pedagogical adaptations or deviations from the model.

1.2. *The DACUM Method*

The DACUM (*Developing A Curriculum*) methodology is based on a pragmatic and collaborative approach to occupational analysis, starting from the premise that expert workers are the professionals best qualified to define their own work. This method posits that any occupation can be accurately described by identifying the specific tasks performed by successful workers, recognizing that each of these actions demands particular knowledge, skills, and attitudes for its effective execution. Under this "learning by doing" philosophy, the process takes place in an intensive workshop where a panel of 5 to 12 experts, under the guidance of a facilitator, breaks down the occupation into general functions and measurable tasks. The final result is a detailed competency map that translates work experience into a structured job profile, which serves as a scientific basis for the design of training programs, intervention monitoring, and performance evaluation systems [14].
The DACUM method offers significant competitive advantages over traditional occupational analysis methods, such as direct observation or individual interviews. One of its main strengths is its time and cost efficiency, as it allows for the consolidation of data in just two days of group work, whereas other methods would require weeks of data collection [14]. Furthermore, by relying on a panel of experts currently performing the work, the method ensures superior content validity, capturing the real behaviors and tasks demanded by the labor market, in contrast to academic approaches that tend to become outdated quickly [9]. Finally, the collaborative nature of the process fosters immediate consensus and a sense of ownership among employers and educational institutions, facilitating a smoother transition from curriculum design to on-the-job implementation [5].

1.3. Analysis of the Nature of the Tasks: Complexity and Difficulty

In the study of human performance, tasks are defined as the fundamental activities that individuals must perform to progress in both their personal and professional lives [11]. The literature has shown that the intrinsic characteristics of these tasks not only condition job performance but also shape an individual's social behavior [11]. One of the most critical variables for classifying these activities is their complexity. Highly complex tasks place substantial demands on the performer's skills, knowledge, cognitive functions, and memory [11]. Therefore, to determine whether performance in an intervention is satisfactory, it is imperative to analyze the relationship between task complexity and the outcome, allowing us to establish whether the observed performance aligns with the expected performance.

Several authors have attempted to define cognitive complexity from multiple dimensions. Wood (1986) proposes a tripartite model [17]:

- **Component Complexity:** Refers to the volume of different acts and information signals that must be processed.
- **Coordinative Complexity:** Linked to the nature of the relationships between inputs and final products.
- **Dynamic Complexity:** Originating from the variability of the environment that alters these relationships.

Other theoretical frameworks have expanded on this view. Campbell (1988) identifies the presence of multiple paths and outcomes, conflicting interdependence, and uncertainty in relationships as sources of complexity [7]. Bonner (1994), for his part, articulates complexity around information processing (input, process, and outcome), emphasizing the quantity and clarity of information [7]. Finally, Harvey and Koubek (2000) extend Wood's work by proposing dimensions of scope, structure, and uncertainty [7]. Taking together, these perspectives reinforce the idea that the quantity, interaction, and variation of the elements of a task are the main determinants of its complexity.

It is essential, however, to distinguish between task complexity and difficulty, concepts often used interchangeably in literature but considered distinct in this research. This paper proposes a critical conceptual distinction: complexity is an inherent property of pedagogical design (the 'what' and 'how' of the educational model of Grupo Ceiba), while difficulty is redefined as operational feasibility ('how frequent'). Thus, while complexity is stable and depends on the design, difficulty is variable and depends on the individual's interaction with their work environment, moving away from the traditional view that subordinates it solely to the subject's perception of self-efficacy.

This distinction between complexity and difficulty underscores the importance of a thorough occupational analysis. In the context of Grupo Ceiba's flexible education model, the facilitator's tasks are highly complex, as the intervention requires not only following a protocol (elements of the educational model) but also adapting information to the student's changing environment (uncertainty). Furthermore, difficulty, understood as feasibility, depends largely on contextual factors.

1.4. The Use of AI to Develop Standardized Instrument Items

The integration of Artificial Intelligence (AI) into the development of items for standardized instruments represents one of the most dynamic frontiers in contemporary psychometrics, as highlighted at the 2026 NCME (National Council on Measurement in Education) conferences. The use of artificial intelligence allows for the automatic generation of items at an unprecedented scale and speed, facilitating the creation of extensive item banks and reducing the risk of content exposure [2;12]. However, recent NCME research emphasizes that this potential must be balanced with strict psychometric rigor; authors such as Hao et al. (2024) and Bulut et al. (2024) warn of the need for "responsible AI" frameworks to mitigate algorithmic bias and ensure construct validity [10; 3]. In this sense, it is proposed that AI should not replace, but rather enhance the work of content experts, using *prompt engineering* and semantic alignment strategies to ensure that the generated items maintain the necessary relevance and fairness in large-scale evaluations [16].

2. Methodology

This research is based on a psychometric instrumental design, aimed at developing and validating an instrument to measure the fidelity of implementation in the Grupo Ceiba educational model. The procedure integrates the DACUM occupational analysis for content delimitation, Haladyna 's (2006) guidelines for scale construction [8], and Shoukri 's (2010) criteria for reliability analysis, item analysis, and internal structure of the instrument [15].

2.1. *Participants*

The study was conducted in collaboration with the Grupo Ceiba Association of Guatemala. Technical validation and piloting were carried out with the 53 Grupo Ceiba facilitators who implement a flexible educational model at their respective locations up to January 2026. Additionally, a panel of experts, comprised of 8 to 12 facilitators with more than 3 years of experience and regional coordinators from the institution, participated in the DACUM workshop.

2.2. *Instrument Development Procedure*

Following the test development cycle proposed by Haladyna (2006), the process was structured in five validation phases [8]:

Phase 1: Domain Definition and Content Validity. To ensure that the instrument accurately represents the essential element of the original design of the educational model, an intensive workshop was conducted using the DACUM (*Developing A Curriculum)* methodology. In this phase, the panel of experts broke down the facilitator's role in general responsibilities and specific, measurable tasks. This process translated the facilitator's role into measurable responsibilities and tasks, ensuring content validity from the design's foundation. Subsequently, these tasks underwent a validation survey to determine their frequency and importance, yielding a criticality index for each. The criticality index is a metric used to rank elements (tasks) based on their frequency of implementation and importance within the educational model [6].

Phase 2: Specification table and item construction. Based on the criticality index, a specification table was developed to determine the number of observation items to be included in the observation instrument, distributed according to the estimated criticality for each one.

Table 1. Table of Specifications

Area of responsibility of the facilitator	**Number of items according to criticality index to be included in the instrument**
Pedagogical agreement	14
Administration	9
Assessment	7
Monitoring	8
Maintenance	1
Divulgation	8
Total	47

Phase 3: Exhaustive generation of possible items to observe in a facilitator. Once the table of specifications was developed, items were generated, worded as observable actions. For this, the artificial intelligence tool Google Gemini was used, and it was asked to suggest all possible ways to observe each task on the task map obtained in the DACUM workshop, without any limit. In the process of requesting the item bank from Gemini, the document containing the educational model of Asociación Grupo Ceiba [1] was provided as part of the prompt. This bank resulted in 316 possible items to include in the questionnaire.

Phase 4: Instrument assembly. From the bank generated in Phase 3, the final items were randomly selected, respecting the proportionality indicated in the specifications table of the previous step (Haladyna, 2006). However, the Grupo Ceiba team that validated the instrument suggested, including only 44 items as three were considered to be redundant. The table shows the total number of items. The instrument is structured using a four-level frequency scale that allows for the assessment of facilitator performance, ranging from "never does it" (1) to "always does it" (4), with intermediate levels of "does it occasionally" (2) and "does it frequently" (3). The final instrument consists of 44 items organized into six job responsibilities (theoretical dimensions), which were directly derived from the key responsibilities identified during the DACUM workshop and further in the table of specification (Table 1).

Phase 5: Piloting the instrument. To strengthen objectivity and mitigate observational bias, a multi-informant evaluation scheme (triangulation) was adopted while piloting the final instrument: (a) the regional coordinator, (b) a student with at least one year of experience, and (c) the facilitator themselves through self-evaluation. The instrument was configured for direct evaluation by the coordinators, based on their frequent visits to the sites. The facilitators responded based on a self-evaluation, and the students based on their experience as students of the facilitator for at least one year.

2.3. Data Analysis

The psychometric quality of the instrument was analyzed under the following criteria:

2.3.1. Item Analysis

An initial exploration was conducted using the frequency distribution of the scale options: "never does it" (1), "does it occasionally" (2), "does it frequently" (3), "always does it" (4) . Subsequently, internal consistency was estimated using Cronbach's alpha coefficient, with the psych package in the R statistical environment.

2.3.2. Internal Structure

To assess construct validity, an exploratory factor analysis (EFA) was performed using the psych package in the R statistical environment. The purpose of this analysis was to verify that the empirical structure of the tasks corresponded to the theoretical dimensions of facilitator responsibilities as defined in the Dacum Workshop and further in the table of specifications (Table 1).

2.3.3. Implementation Fidelity Exploration

Finally, the facilitators' overall scores were estimated, disaggregated by the three profiles of actors participating in the study, to compare the reported implementation levels.

3. Results of the Instrument Validation

3.1. Item Analysis

Analysis of the response frequencies reveals a marked tendency towards the higher categories on the scale: **"I always do it" (4)** and **"I do it frequently" (3).** The category **"I do it occasionally" (2)** has a significantly lower frequency, while the option **"I never do it" (1)** is selected marginally.
This distribution suggests that, while the four-level scale effectively discriminates the degree of implementation among those already applying the model, it has a limited capacity to identify those who have not implemented it at all (ceiling effect). Consequently, calculating the implementation fidelity score is the proportion of frequency of implementation over complete implementation of the task.

Table 2. Average proportion of selection of each point on the scale

	Number of items	I never do it	I do it occasionally	I do it frequently	I always do it
Pedagogical agreement	14		4.6363	44.7357	47.8429
Administration	8	1.9	5.6857	37.5222	54.0889
Assessment	7		3.04	45.2857	48.7857
Monitoring	7	6.3	7.5571	47.9857	38
Maintenance	1		1.9	20.8	73.6
Divulgation	6	1.9	6.9333	39.3	49.6667

The 44-item scale exhibits excellent internal consistency, with an overall Cronbach's alpha coefficient of 0.96 [0.95, 0.96]. Item homogeneity is adequate, with an average inter-item correlation of 0.33. The reliability analysis if one item is removed (Reliability if an item es The deleted items demonstrate that excluding any item would not significantly increase the overall coefficient, suggesting that all items contribute positively to the consistency of the construct. Furthermore, all corrected item-total correlations exceed 0.40, confirming the strong discriminatory power of each item within the scale.

3.2. Factor Analysis

Exploratory factor analysis confirms the existence of a robust general implementation factor underlying the entire instrument (MR1). However, specific dimensions are identified that group teachers' practices around pedagogical agreements and administrative practices (MR3); items from sections A (A_2_12_D,

0.648) and B (B_4_32_D, 0.551) load here. Monitoring and administrative tasks (MR4) are also included; items from section D, such as D_2_11_D (0.658) and B_1_6_D (0.624), load heavily here. Finally, evaluation (MR2) is defined by items C_1_6_D (0.600) and C_3_24_D (0.501). The presence of cross-loadings suggests that the Grupo Ceiba educational model functions as an integrated system where the components mutually reinforce each other. This finding is consistent with the literature on the dynamic complexity of work tasks.

Table 3. Factor Loads

Factorial Dimension	Load	Description of (MR)
MR1: General Implementation	Dominant	Robust factor that underlies the entire instrument.
MR3: Pedagogical Agreement	0.648	Pedagogical agreement and administrative tasks.
MR4: Monitoring	0.658	Monitoring and administrative tasks.
MR2: Evaluation	0.600	Evaluation tasks.

3.3. Implementation Fidelity

To estimate the implementation fidelity results, fidelity of implementation of the facilitators scores as evaluated by three people were obtained: 1) the coordinator (immediate supervisor), 2) their students, and 3) themselves in a self-evaluation format.

With 95% confidence, the implementation fidelity according to the facilitators' self-assessment falls between CI (0.83, 0.88). With the same confidence level, the students find a slightly higher implementation fidelity CI (0.88, 0.92). Finally, the coordinators find a level of implementation fidelity very close to the facilitators' self-assessment CI (0.81, 0.87). The graph shows that the coordinators' evaluations of the facilitators vary more than the self-assessments and the students' evaluations. Presumably, this is because the coordinators can compare the work of several facilitators under their supervision. In all three cases, implementation of fidelity exceeds 80%.

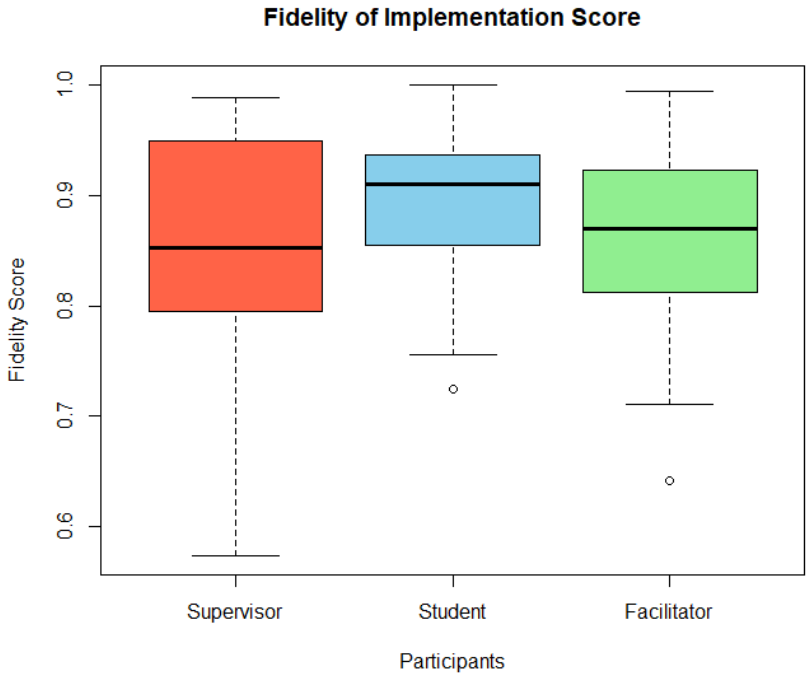

Fig. 1. Average score per group of participants

4. Conclusions

The validation process for a monitoring instrument, initially based on task analysis using the DACUM methodology, ensures that the measurement will focus on the specific operational aspects of the intervention. In this sense, designing tools with a focus on faithful implementation is essential for project sustainability, as it enables informed decision-making and continuous, evidence-based improvement.
Under this rigorous methodology, the analyses performed confirm that the resulting monitoring instrument is reliable and valid, meeting the necessary psychometric properties to ensure that the fidelity of its implementation is based on precise measurements. In the item construction phase, the use of Artificial Intelligence emerges as a valuable tool for generating a comprehensive item bank; however, its application must be strictly subject to the model's normative documents and undergo qualitative validation with the intervention's leadership to ensure its relevance.
Finally, the results of this study suggest that, discriminating againstw response frequency in certain categories, it is essential to preserve a scale that includes at least three levels of performance: one that documents the absence of implementation and two that allow discriminating the gradation of the model's presence in the work context.

REFERENCES

[1] Asociación Grupo Ceiba. (2025). *Modelo educativo de educación extraescolar Asociación Grupo Ceiba*. GPEKIX. https://gpekix.org/recurso/modelo-educativo-de-educacion-extraescolar-asociacion-grupo-ceiba/
[2] Bißantz, J., et al. (2024). *Applications of generative AI in large-scale assessment development*. Proceedings of the NCME Annual Meeting.
[3] Bulut, O., et al. (2024). Guidelines for responsible AI in educational measurement. *Educational Measurement: Issues and Practice*.
[4] Century, J., Rudnick, M., & Freeman, C. (2010). A framework for measuring fidelity of implementation. *American Journal of Evaluation, 31*(2), 199–218.
[5] Crockett, J. B. (2012). Developing a Curriculum (DACUM): A process for analyzing the role of the special education administrator. *Journal of Special Education Leadership*.
[6] Dierdorff, E. C., & Wilson, M. A. (2003). A meta-analysis of job analysis reliability. *Journal of Applied Psychology*, *88*(4), 635–646.

[7] Gage, N., MacSuga-Gage, A., & Detrich, R. (2020). *Fidelity of implementation in educational research and practice*. Wing Institute.
[8] Haladyna, T. M. (2006). *Handbook of test development*. Lawrence Erlbaum Associates.
[9] Halasz, I. M. (1994). *The DACUM process*. Center on Education and Training for Employment, Ohio State University.
[10] Hao, J., et al. (2024). *The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges*. WSU Research Exchange.
[11] Liu, P., & Li, Z. (2011). Task complexity: A review. *Reviews of Human Factors and Ergonomics*.
[12] Luecht, R. M. (2025). *Automated item generation and the future of test construction*. NCME/AIME-Con Proceedings.
[13] McLean, R., & Gargani, J. (2019). *Scaling impact: Innovation for the public good* (1st ed.). Routledge. https://doi.org/10.4324/9780429468025
[14] Norton, R. E., & Moser, J. (2008). *DACUM handbook*. Ohio State University, Center on Education and Training for Employment.
[15] Shoukri, M. M. (2010). *Measures of interobserver agreement and reliability* (2nd ed.). CRC Press.
[16] Suárez-Álvarez, J., et al. (2024). Using artificial intelligence in test construction: A practical guide. *Psicothema*.
[17] Wood, R. E. (1986). Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes, 37*(1), 60–82.