Revista de Lenguas Modernas, N.° 40, 2025 / 01-21
ISSN electrónico: 2215-5643
ISSN impreso: 1659-1933
DOI: 10.15517/Y7X41780

Cognitive Task Analysis as a Tool to Enhance Raters' Reliability at Scoring a Writing Admission Test

El Análisis Cognitivo de Tareas como Herramienta para mejorar la Confiabilidad entre Evaluadores al Calificar una Prueba de Admisión de Escritura

Mag. Jorge Calvo Jiménez
Universidad de Costa Rica, San José, Costa Rica
Ministerio de Educación Pública, San José, Costa Rica
jorge.calvojimenez@ucr.ac.cr
ORCID: 0009-0007-9876-8665

Mag. Ana Carolina González Ramírez
Universidad, de Costa Rica, San José, Costa Rica
ana.gonzalezramirez@ucr.ac.cr
ORCID: 0000-0003-1937-6465

Abstract
Cognitive Task Analysis (CTA) has proven useful to design training programs in diverse areas of human professional and academic endeavors. This paper explores its use as a tool to increase levels of validity, reliability, and fairness, among other assessment components, and therefore, project a forthcoming calibration of the rating process of a high-stake writing evaluation, by analyzing the thoughts and process description of expert raters as they graded candidates' essays. This research implemented CTA data elicitation techniques in the form of recordings, transcripts, and reports provided by subject matter experts (SMEs). First, in stage one, volunteer students were asked to write an essay that had the same input, purpose, and conditions of the real test. Next, four SMEs graded three of these essays and recorded their thoughts in oral and written form based on specific guidelines. The information was analyzed, and six preliminary categories of analysis were identified. In stage two, after a qualitative analysis and systematization of information, three of the six categories were determined to be part of the raters'

Recepción: 28-04-2024 Aceptación: 08-07-25

scope of action: work operationalization (procedural knowledge), cognitive processes for the rating tasks (declarative knowledge), and the assessment instrument. During this phase, the test takers wrote real essays. Two raters were assigned three specific pieces of writing from this batch and asked to record their thoughts and process as they performed the task. The results of the analysis of the three categories are expected to prove useful in addressing the improvement of internal processes in the long term by including them in future rater training processes.

Keywords: Cognitive Task Analysis, writing task, writing rating, task operationalization, validity

Resumen

El Análisis Cognitivo de Tareas (CTA) ha probado ser útil para el diseño de programas de entrenamiento en diversas áreas del quehacer profesional y académico del ser humano. Este artículo explora su uso como una herramienta que incremente niveles de validez, confiabilidad y justicia, entre otros componentes evaluativos, y de esta manera prever su uso en una futura calibración del proceso de jueceo de una evaluación de escritura de altas consecuencias al analizar los pensamientos y descripción de procesos de jueces expertos mientras estos calificaban ensayos de candidatos. Esta investigación implementó técnicas de obtención de datos del CTA tales como grabaciones, transcripciones, y reportes provistos por expertos en la materia (SMEs). Primeramente, en la etapa uno, se les solicitó a estudiantes voluntarios escribir un ensayo con el mismo recurso, propósito y condiciones del examen real. Luego, cuatro SMEs calificaron tres de estos ensayos y grabaron sus pensamientos en forma oral y escrita basados en indicaciones específicas provistas por los investigadores. La información fue analizada y seis categorías de análisis fueron identificadas. En la etapa dos, tras un análisis cualitativo y la sistematización de información, se reconocieron tres de las seis categorías como elementos clave del campo de acción de los jueces: operacionalización del trabajo (conocimiento procedimental), procesos cognitivos para las tareas de jueceo (conocimiento declarativo), y el instrumento de evaluación. Durante esta fase, los ensayos escritos por los candidatos fueron reales. A dos jueces se les asignó tres composiciones específicas, y se les solicitó grabar sus pensamientos y procesos mientras realizaban la tarea. Se espera que los resultados del análisis de las tres categorías sean útiles para el mejoramiento de procesos internos a largo plazo y se refieran en futuros entrenamientos para jueces.

Palabras clave: análisis de tareas, procesos cognitivos, evaluación escritura, operacionalización de tareas, validación

Introduction

This paper analyzes the information gathered through the implementation of a five-stage Cognitive Task Analysis (CTA) research to collect insightful information from Subject Matter Experts (SMEs) to inform a future calibration program for writing raters. Although CTA has been used in the areas of education and training development, according to the literature consulted, because of the cognitive demands placed upon subjects and the resource-consuming nature of the method, the method does not seem to have been commonly used in rater-calibration processes in the assessment of linguistic tasks. It has not been used to validate assessment instruments and tasks either. Given the valuable expert insights that CTA can provide, its potential as a tool to improve internal processes, rating performance, and even scoring tools must be harnessed.

This study aimed at answering research questions that dealt with the implementation of CTA during a scoring process of essays, the theoretical and operation foundations that would support such usage, and the conditions necessary for the execution of the methodology. While rating some pilot test essays, four SMEs recorded their insights in the first stage of the study. Based on the data obtained, six categories of input were identified, from which three were selected to be further investigated in the second stage. The results show the relevance of the areas chosen in terms of their influence on the raters' performance. Future goals to pursue include the design of a rater training program based on the categories inferred and the creation of rating protocols based on it.

Background

The Master's Program in Teaching English (MA in TEFL or *Programa de Posgrado en Enseñanza del Inglés como Lengua Extranjera*, PPEILE), at the University of Costa Rica (UCR), is a post-graduate program that was born in 1998 and which has become a highly successful language education program in the field of English teaching (Solís Hernández, 2009) in terms of graduates' degree of satisfaction. Designed as a two-year program (60 credits), because of the advanced content and language courses that it includes, such as SLA, Applied Linguistics and skill-integration courses, and students' multiple responsibilities, most people complete the program in three years. PPEILE also requires candidates to hold a BA in English or a related field, a Grade Point Average (GPA) of 8 or higher, and "a grade of 8 or above on an English exam that assesses speaking and writing skills" (Solís Hernández, 2009). As observed, the master's program in TEFL is a well-organized program that requires a strong commitment from its students since the beginning of the endeavor.

To increase the possibility of having those people first enrolled finish the program successfully, PPEILE requires entry candidates to master the English language both in its oral and written form, which means that their English performance should be at the C1 level according to the Common European Framework of Reference for Languages (CEFR). The Council of Europe (2001) establishes that a person granted a C1 level possesses an "effective operational proficiency" and shows "good access to a broad range of

language, [fluency], spontaneous communication… [and] a fluent repertoire of discourse functions" (p. 32). This person can also "produce clear, smoothly flowing, well-structured speech, showing controlled use of organizational patterns, connectors and cohesive devices" (p. 32). However, there are registration candidates who may be either unaware of this requisite or whose language skills are below those needed to undergo the program successfully.

Therefore, PPEILE resolved to have an admittance exam that comprises two core components: a written part (essay) and an oral interview. The responsibility for designing, administering, and grading the essay-writing component of the examination was handed to the Programa de Evaluación en Lengua Extranjera (PELEx) in 2020 as this department has specialized in assessment development and administration. In recent years, researchers at PELEx have been working in the development of measurement and evaluation tools whose results are meant to be fair and valid for users. The growing responsibilities of PELEx include the design, administration, and evaluation of different language proficiency tests in diverse languages and which address various purposes and stakeholders. The endeavor implies following the strictest protocols and guidelines to make sure their assessing tools fulfill quality requirements.

Rationale

Currently, legitimate evaluation practices in education, and especially in the assessment of language proficiency, require entities in charge of providing assessment services to be transparent and ensure fairness and validity for each of their products. Messick explains the continuous search of these principles by stating that "validity, reliability, comparability, and fairness are not just measurement principles, they are social values that have meaning and force outside of measurement whenever evaluative judgments and decisions are made" (Messick, 1995, p. 742). Therefore, entities involved in assessment should hold these evaluation principles close to their everyday practices. PELEx is aware of adhering to the principles outlined above and is committed to meeting them.

The admission test for PPEILE, now in charge of PELEx, is the candidates' door to being admitted to a life-changing opportunity: that of pursuing a post-graduate degree. Precisely because of the high consequences that passing or failing this evaluation has for candidates, researchers at PELEx intend to calibrate the performance of the raters when scoring candidates' written output by using Cognitive Task Analysis (CTA), which is a set of methods applied to elicit the cognitive steps necessary to correctly perform a task. This article will describe the chosen methodology, as well as its respective elements and operationalization.

Not only would this research method prove useful in shedding light as to the thinking processes that experienced writer raters have, but it would also provide valuable information about other areas and elements of the test that need to be scrutinized for validity purposes, such as reviews of the assessment construct, its validity, reliability, generalizability, comparability, and fairness, as well as the usefulness

and assessment scoring tool used by raters. By analyzing raters' processes as they rate essays, it could also be possible to identify and eventually address performance issues that some could believe threaten reliability, for example, uncertainty, which even experienced raters admit experiencing as they carry out a rating task. Although "uncertainty was not found to vary as a function of rater consistency or severity/leniency" (Honko et al., 2023, p. 11), identifying its possible causes and addressing those 'external' ones that can be modified and improved could enhance raters' self-confidence and make their work less burdensome, especially in high stake examinations, where "raters know the test scores will have major consequences for the examinees" (Aloha, 2016 in Honko et al., 2023, p. 3). Working on raters' confidence, then, can improve test reliability.

The data elicited by CTA can also help bridge the information gap that occurs when experts train others. As stated by Feldon et al. (2006, as cited in Lyness et al., 2021), "capturing unobservable thoughts, decisions, and judgments from multiple experts is required as studies show that experts may omit up to 70% of the critical information when describing how to replicate expert performance to others" (p. 12). By collecting these insights through think-aloud recordings, reports, and interviews, such relevant information surfaces as key data that cannot only be analyzed by researchers but that can also be used by future raters in future calibration programs.

Review of the Literature

Cognitive Task Analysis (CTA)

The concept of Cognitive Task Analysis (CTA) emerged in the 1970's and it derives from former approaches such as Task Analysis. The first printed mentions of CTA appeared in J. Gallagher's work "Cognitive/information processing psychology and instruction: Reviewing recent theory and practice" (1979), in which the author relates the term to the fields of production systems and artificial intelligence. Since then, CTA has evolved to become a set of methodologies that allows the elicitation of knowledge about several work and research areas, as well as knowledge of domains, scenarios, procedures, routine plans and goals, and reasoning, among others.

The transition towards a *study* methodology that included more theoretical and strategic components is reflected in the work of Schraagen et al. (2000), who define CTA as "the extension of traditional task analysis techniques to yield information about the knowledge, thought processes, and goal structures that underlie observable task performance" (p. 3). Over the years, supporters of CTA have focused their work on eliciting the cognitive processes and procedural knowledge that underlie experts' task performance, especially in dynamic and professional domains.

Klein and Militello define CTA as "a set of methods to elicit, explain, and represent the mental processes involved in performing a task" (2001, p. 168). This concept of CTA is an all-encompassing approach whose main purpose is to bring forth the way the

mind works, how it leads people to decision-making and reasoning, and how information is processed; indeed, it portrays a scope that covers theory and operationalization of the approach. Another conceptualization that shares this view is that of Knisely et al. (2021), which states that CTA "focuses on the underpinning mental framework, thought processes, and knowledge behind the performance of a task" and that it "can be used to identify hidden and ineffective cognitive strategies as well as tasks that induce high cognitive demand" (p. 2).

Because of the broad coverage that CTA offers to research and because of the different methodologies attached to it, a more suitable definition to fulfill the purposes of this article would say that CTA is a multi-methodological approach that is used mainly to outline the mental processes, cognitive elements, reasoning and operationalizations necessary to perform tasks in the field of language proficiency assessment. Eventually, this understanding of CTA could contribute to the design of training for raters of academic essays, with a high level of proficiency.

Applications in Assessment

The results of CTA can be used in process analysis, training design, improvement of work systems, task protocols and workflows, and even assessment. This approach sets the principles required to better comprehend processes, to construct the instruments through which researchers can elicit and collect information, as well as to analyze processes and products. Hoffman and Militello (2008) summarize this by stating that:

> The power of CTA is in how the needs and historical trends have come together to be applied to the challenges of the information age: using qualitative and quantitative methods to study cognition and expertise in the context of work, to explore a work domain with the objective of uncovering cognitive complexity, and to apply this knowledge to the design of tools, technologies, and work systems. (p. 8)

Early applications of CTA took place in the fields of computer system interfaces and military applications and have been expanded to areas as diverse as cognitive science (Clark et al., 2008). Hoffman and Militello (2008) include uses in other areas such as aircraft tasks, industrial process control, medicine proctoring, and electronics troubleshooting, which can also benefit from CTA (p. 6). The authors add other uses of the approach, for example, the design of automated decision aids, interfaces and workstations, the elicitation of expert knowledge for intelligent systems, the preservation of corporate knowledge, and error identification and mitigation.

Education is also an availed field of CTA methods, as it provides "general-purpose tools for conducting inquiries and learning from SMEs [subject matter experts]" (Crandall et al., 2006, p. 252). The scope of application of the approach would reach common tasks of the education system and involves any process of the system, such as assessment. In addition, Otálora (2019) highlights more advantages to the utilization of CTA in tasks of educational assessment by pointing out two fundamental aspects: the intended learning

of the assessment and the symbolic framework, which play a role in the understanding of the tasks (p. 5). Consequently, CTA implies the socialization and understanding of experts' knowledge within assessment which play a key role in the assurance of validity.

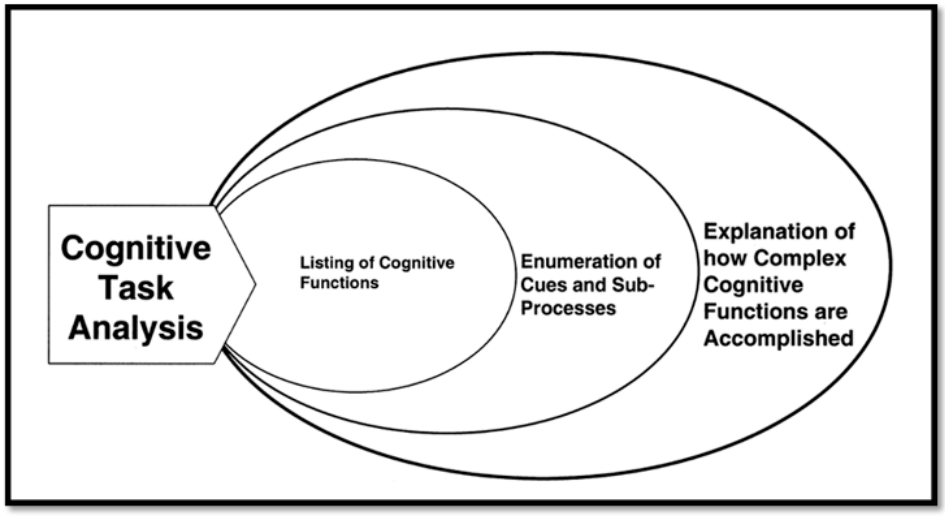
Key Elements of CTA

1. Subject Matter Experts (SMEs). CTA techniques elicit knowledge and reasoning behind process execution from subject matter experts (SMEs); this data is then analyzed to derive valuable representations. In simple words, SMEs are professionals who have extensive experience that permits them to succeed rapidly and consistently at a class of tasks (Clark et al., 2008). In "Using cognitive interviews to validate an interpretive argument for the ETS iSkills™ assessment", by Snow and Katz (2009), SMEs are described as individuals who possess specialized expertise and knowledge within a particular domain pertinent to the ETS iSkills™ assessment. SMEs play a crucial role in ensuring the validity and relevance of the assessment items through their discerning insights and domain-specific knowledge.

SMEs also refine the clarity and relevance of assessment items and bolster the integrity of the assessment process. Indeed, Snow and Katz (2009) claim that SMEs offer invaluable contributions to the validation process because they "provide essential feedback on the relevance and clarity of assessment items, thereby enhancing the overall validity of the assessment" (p. 123). The worth of involving SMEs in assessment systems and processes is well explained in *Standards for Educational and Psychological Testing* (AERA et al., 2014), as it explains that an SME's judgements can assess the importance, criticality, and/or frequency of tasks, and that these professionals are able to elaborate on performance-level or achievement level descriptors and can also determine the role of cognitive ability in a performance domain.

2. Outcomes. The interpretation of the expected outcomes can be more accurate if these are foreseen and predicted since the construction of the validity argument. Klein and Militello (2001) state that the outcomes obtained from CTA studies might range from the enumeration of cognitive functions, through the dissection of these functions, and up to the generation of an explanation of how the function is accomplished. They created a scheme (Figure 1) to show how CTA studies can target different outcomes, and that depending on the expected outcome, the scope of what is being explained and discovered can increase or decrease.

Figure 1
Range of Outcomes for CTA Projects



Note. Figure taken from Klein & Militello, 2001, p. 172

In Figure 1, Klein and Militello (2001) propose that in research in which a basic CTA approach is used, the first scope of outcomes is the listing of cognitive functions required to complete a task in the form of judgments and decisions. The second group of outcomes, which includes cues and sub processes, might be useful to highlight the patterns and strategies that go into the judgments of the task performance. Lastly, the largest scope of possible outcomes contains insights regarding the way cognitive functions and tasks are performed. No doubt a high-stakes assessment, as most of the ones administered by PELEx, would benefit from CTA processes whose target is reaching the widest scope of outcomes.

3. Declarative and Procedural Knowledge. As stated, SMEs are a source of knowledge and expertise which allows the implementation of CTA methodologies and processes. The types of knowledge that they offer to CTA processes are declarative and procedural, which interact in the acquisition of competences or skills; this is a process that "typically commences with the procurement of declarative knowledge, which is used to compile procedural knowledge for using a specific skill or exhibiting a desired behavior" (McManus, 2024, p. 149). In fact, McManus (2024) specifies that the former "reflects one's conceptual understanding of something", whereas the latter "pertains to the way a certain behavior is performed" (p. 149). Consequently, declarative knowledge refers to what is needed to be *known* to perform a task, while procedural knowledge deals with what must be *done* to fulfill that task.

Declarative knowledge implies cognition related to specific academic or professional fields, which includes principles, theories, taxonomies, methods, techniques, type and amount of information, and knowledge of networks. It is neurologically stored as

"hierarchically structured propositional, episodic, visuospatial information that is accessible in long-term memory and consciously observable in working memory" (Clark et al., 2008, p. 8). The importance of identifying and working with this type of knowledge resides on the fact that it provides people with conceptual understanding of processes and frameworks of schema-based representations to analyze complex problems, retain and recall information, and to solve novel and complex problems (Clark et al., 2008, p. 9).

Hoffman and Militello (2008) assert that procedural knowledge refers to procedures, rules, strategies, and, of course, knowledge about what to do when performing tasks, or applying good judgment in unique situations (p. 151). In order to have a fortified and robust cognition to proceed, a person first requires a solid declarative knowledge of the theory and principles, which will follow them to operationalize them accurately. This transition from one type of knowledge to another can be explained as "reasoning or knowledge [that] originates as an analytic, conscious, deliberative, stepwise process [declarative] and evolves into rapid, automatic, nonconscious, understanding or immediate perceptual judgments [procedural]" (Hoffman & Militello, 2008, p. 154).

4. Critical Decision Method (CDM). According to Hoffman and Militello (2008, p. 181), Klein, Calderwood, and MacGregor are the theorists behind the formulation of the term *Critical Decision Method*. The authors present their method as one which uses "a set of cognitive probes to determine the bases for situation assessment and decision making during nonroutine incidents" (Klein et al., 1989, p. 462). This method, when associated to CTA, can elicit information about how experts solve the different problems that come out while performing a task. Hoffman and Militello (2008) state that CDM can be summarized as a type of interview that involves task retrospection.

Watkins and Visser (2009) explain the method as the CTA researcher utilizing SME's information recalled from a critical or uncommon situation of a particular field or area, and then both working together to systematically identify declarative and procedural knowledge within the information given. Hoffman and Militello (2008) indicate that the interview in which both professionals discuss those moments must be comprised of questions addressed to:

> Elicit information that is specific and meaningful: strategies and the basis for decisions, and the perceptual cues on which the decision maker relies on types of information that were not ordinarily the focus in either laboratory research on expertise or applied knowledge elicitation projects. (p. 182)

Klein et al. (1989) add on this idea by explaining that the questions must provoke deeper responses, which "require the decisionmakers to reflect on their own strategies and bases for decisions," and that this "makes the method so appropriate to a variety of knowledge elicitation needs" (p. 462). Hoffman et al. (1998) justify the application of CDM in CTA because the latter "is used more widely in the elicitation, preservation, and

dissemination of expert knowledge and … more widely as the basis for the design of complex cognitive system", which are discussion issues that "become increasingly critical" (p. 254).

The Five Common Stages of CTA Methodologies

Cognitive Task Analysis has become an approach comprised of several methods and tools that has undergone some modifications according to specified purposes, such as Applied Cognitive Task Analysis, or ACTA (Militello & Hutton, 1998). Therefore, its operationalization level of difficulty also varies, as it may be extremely simple or complex to apply. The same goes for the preparation of the materials and planning needed for the execution. Moreover, despite the specific methodology chosen within the approach, CTA will always allow both the recognition of processes that are not readily observable and the gathering of detailed information during an abbreviated period.

Nonetheless, before deciding on what methodology to apply for a CTA project, researchers must determine which upper principles every methodology must follow. Klein and Militello (2001) point out three principles or criteria: discovery, communication, and impact. First, in terms of the discoveries, the most exciting in a CTA study are those that result in an explanation or insight regarding the way a cognitive function is performed. Second, successful communication occurs when the CTA team provides the potential user of the findings with an understanding that is sufficiently vivid so that the user can take the cognitive processes into account in designing an intervention. Finally, impact has to do with findings and results of the CTA process being "put into action" (Klein & Militello, 2001, p. 176).

On a similar focus, Crandall et al. (2006) state three critical aspects that would bring success to any CTA project: "knowledge elicitation, data analysis, and knowledge representation" (p. 2). These more complex and theorized concepts relate to the methodology chosen for the CTA applied in this research project.

Clark et al. (2008) have estimated that CTA can be driven by more than one hundred known methodologies and their respective several derivations. It is necessary to recognize that every single research project or assessment process would need specific methods and strategies. Hoffman and Militello (2008) share Wei and Salvendy's four broad families of CTA methods, each of which is comprised of several similar methods. The work is shown in Figure 2, which includes a column for the most widely used methodologies of each family, and another column for appropriate uses. This type of webs of CTA methods seems common among theoreticians of the field and allows the codification of strengths, weaknesses, and appropriate uses of all methods.

Figure 2

Major Families of CTA Methods

Family	Particular CTA Methods	Appropriate Uses
Observations and interviews	Many types of structured interviews are listed, such as the Critical Decision Method	Useful to define and circumscribe the domain. Useful for domains where specific task procedures are not well definedUseful for the analysis of tasks that are skill based
Process tracing methods	Verbal reports	Useful when it is easy to define representative tasks and scenariosUseful when it is important to evaluate task (or dual-task) performance. Useful for the analysis of tasks that are skill based or rule based
Conceptual techniques	Graphing tasks, ratings tasks	Useful when it is important to reveal domain knowledge. Useful for the analysis of tasks that are rule based or knowledge based
Formal models	ACT-R, GOMS modeling	Useful for modeling tasks that do not change much. Useful for the analysis of tasks that are rule based or knowledge based

Note: ACT-R = Adaptive Control of Thought-Rational; GOMS = Goals-Operations-Methods-Selection Rules.

Note. Taken from Wei and Salvendy (2014) in Hoffman and Militello (2008, p. 64).

Tofel-Grehl and Feldon (2013) implemented what they called the *Process Tracing Method* to delve into experts' cognition and elicit procedural directions that support effective task performance. This method adheres to the claim of Schraagen et al.; Coffey and Hoffman; Cooke; Crandall et al.; Hoffman et al.; and Clark et al.; all cited by Knisely et al. (2021), regarding the most common CTA methodologies being comprised of five stages, which were followed to accomplish the present study. Table 1 shows the five stages along with a brief description for each of them.

Table 1

The five stages of CTA methodologies.

Stage	Description
1) Collect preliminary knowledge.	Researchers identify the target performance goals, and get to know the task domain, general knowledge, and processes.
2) Identify knowledge representation.	SMEs describe procedural knowledge to fulfill what is expected.
3) Apply methods of obtaining focused knowledge.	SMEs describe procedural knowledge and declarative knowledge necessary to perform the tasks.

4) Analyze and verify the data.	Researchers create a classifying system to split the elicited knowledge into categories and verify it for accuracy and completeness.
5) Give format to the results for the intended application.	Researchers create formats and schemes for the resulting declarative and procedural knowledge, specifically designed for training, decision-making processes, and tutorials.

Note. Adapted from Knisely et al. (2021).

Given the widespread use of these five stages among researchers and field studies, the present study replicates this CTA process, which contributes to a better understanding of the cognitive processes involved in every stage and also provides opportunities to improve the performance of future raters as it can be used to inform new training programs that would contribute to the reliability of the test and its results. CTA also helps to uncover SME's expertise in an immediate manner that avoids retrieval issues or data filtering by asking them to describe what they do as they are doing it.

Previous research studies

During the process of constructing a solid theoretical and referential framework to implement Cognitive Task Analysis (CTA) in the scoring process of the academic compositions in the entrance examination for PPEILE, reference works with similar goals and focused on assessment of the writing competence of a second language were looked for. Nonetheless, only a couple of studies were found with such a specific topic. Therefore, it was necessary to expand the search criteria and look for other studies on education and training in different areas.

In his 2011 paper, Barkaoui explores the effects that think-aloud protocols may have on essay raters' performance. The conclusions of his qualitative analysis showed that thinking-aloud protocols, a key component of this CTA study, do affect the operationalization and rating severity of some raters. As they verbalized their thoughts, raters were drawn to detect micro-level problems (spelling, grammar) but obscured macro-level comprehension of the texts.

Cannon-Bowers et al. (2013) propose the use of CTA to develop simulation-based training protocols for two specific combat casualty care tasks: cricothyroidotomy (incision in the cricothyroid membrane to establish an airway for oxygenation and ventilation) and hemorrhage control. The results of their work show that CTA is useful to fulfill the training objective and that it allows the generation of usable data to feed a simulation-based training system designed for the Department of Defense of the United States of America. This work provides the PEL-Ex-CTA project with a clear example of the technical steps and methods that should be addressed to establish an institutionalized training program for raters of the written component of the master's test.

An academic work, more approximate to the assessment of learning outcomes, which entails the intention of the master's program at UCR, is the one portrayed by Burgos et al. (2022), who present a CTA addressed to prompt the knowledge and abilities of more than one hundred preservice primary education teachers to interpret students' responses to probability comparison tasks. The researchers aimed to identify the use of incorrect strategies used and proportional reasoning in mathematical activity. The cognitive analysis of the subjects' performance was intended to go beyond common tasks in the classroom and help to establish diverse ways of institutionalizing mathematical knowledge. The elicitation these future teachers' thinking processes while performing the task evidenced their mathematical declarative and procedural knowledge, or lack of both, which will foster actions at the university to prepare preservice teachers to adequately guide their students to solve cognitive mathematical tasks in schools. The current research study shares the same goal: finding declarative and procedural knowledge accurate enough to let raters of a written task perform their responsibility with fairness and in a technically correct manner.

In the article "Cognitive Task Analysis as a Methodological Strategy for Understanding and Explaining Human Cognition" by Otálora (2019), the author explores the utility of CTA for deconstructing complex cognitive tasks into their component processes. Otálora highlights the role of CTA in informing the design of educational training processes by facilitating the analysis of the task at hand and deconstructing its cognitive demands. CTA, then, proves useful as a source of insightful data in problem solving tasks that can be used to support the grading process by breaking it into its essential components.

In their paper "Cognitive Task Analysis-based training: a meta-analysis of studies," Tofel-Grehl and Feldon (2013) concentrate on examining the use of CTA as a valuable approach for analyzing the components of effective, professional task performance. Although helpful in task articulation and in developing instructional design, the authors acknowledge that little research has been done in its value for training. The present research study aims at addressing such a gap in the literature by delving into the use of CTA in scoring training processes.

Research Questions

The following research questions guided this study:

1. How to implement a Cognitive Task Analysis for the scoring process of the academic compositions in the Admission Test for the master's degree in English Teaching as a Second Language, aimed to create a proper calibration of the raters?
2. What are the theoretical and operational foundations to determine a suitable methodology to implement a Cognitive Task Analysis for the scoring process of the academic compositions in the Admission Test for the master's degree in English Teaching as a Second Language?
3. What is needed to execute the phases of a Cognitive Task Analysis for the scoring process of the academic

compositions in the Admission Test for the master's degree in English Teaching as a Second Language, to create a proper calibration of the raters?

Methodology & Discussion

The current study used a qualitative methodology that focused on the analysis of information gathered from primary sources (SMEs) who shared their knowledge and expertise as they performed a specialized rating task of essays written for an entrance exam. It consisted of a pilot test, a first stage of analysis with four subjects, and a second stage with two subjects. Raters' verbalizations were recorded by them, transcribed, and later analyzed. All these experts rated the essays remotely. Participation in the study was voluntary.

Stage 1

The written component of the entrance exam for PPEILE was designed based on the specifications provided by the board that leads the program. PELEx was given examples of previous tests that helped to identify what the MA program looked for in terms of format, administration, and grading. In the case of the latter, a new evaluation rubric was created based on the specifications and the exam objectives shared by the MA in TEFL program. Once the new exam and the evaluation rubric were approved by the PPEILE board, a pilot exam was developed to assess the feasibility of the administration protocols, the access to the online platform where the test would take place, the validity of the test questions and materials, and the use and assessment of the grading guidelines that were developed by PELEx.

This first stage involved the rating of five essays written by the same number of volunteer students from the BA in English in a pilot implementation of the test. The students were undergoing the last semester of their major, and they did not intend to register in the PPEILE. These five volunteers took the test in a controlled, supervised environment where they were given all the necessary materials to carry out the task in similar conditions to those of the actual administration of the test; to safeguard their anonymity, false usernames and passwords were used to log in the online platform of the test.

Once the essays had been collected by PELEx, these written pieces were all sent to the raters who took part in this stage of the study. Each rater received three essays to analyze, whose distribution can be seen in the table below.

Table 2
Distribution of essays to grade by raters

	Essay A	Essay B	Essay C	Essay D	Essay E
SME 1	x	x			x
SME 2	x	x			x
SME 3			x	x	x
SME 4			x	x	x

As seen in Table 2 above, SMEs were paired up and given the same essays to rate. It is worth highlighting that all raters analyzed Essay E, which served as a point of reference.

During this first stage, SMEs were asked to record themselves as they were grading the three essays assigned. Once raters had finished grading the essays, they were also asked to further elaborate on the thoughts and ideas expressed in the audio by listening to their recordings and writing a report-like document in which they explained and expanded on aspects they thought were relevant in their rating process. One of the researchers analyzed the recordings and the report provided by each SME; a set of questions was created to delve into details and further explanations the researchers deemed relevant to fulfill the objectives of this study. Finally, each SME was interviewed by one of the researchers with the instrument created (questions) to elicit more detailed information of their thoughts and specific cognitive processes while performing the scoring tasks. The last step was the systematization of the information collected in the semi-informal interviews with each rater. Based on the data collected from the audios, written reports, and interviews, six general, preliminary categories of information were found to be common among all raters; these six categories are detailed in the Results section below.

Stage 2

Because of internal processes at the institution and to minimize the number of raters involved in the actual analysis and rating of the essays for PPEILE, only two raters could grade the candidate's essays required in the entrance exam: SME 2 and SME 3. This decision was grounded on aspects such as the raters' availability by the time Stage 2 of the study was to be carried out and the small difference regarding the number of the points allotted by them to Essay E based on the evaluation rubric, as the table below shows.

Table 3
Total number of points awarded to Essay E by raters.

	SME 1	SME 2	SME 3	SME 4
Points allotted	13	11	12	9

Note. The maximum number of points that an essay could obtain was 13 based on the evaluation rubric.

The second stage involved the rating of the entrance exam essays by the two raters. There were 13 compositions to grade, as this stage involved the actual rating of PPEILE candidates' essays. This second stage asked the SMEs to carry out another "think-aloud" exercise in which they should record themselves as they scored three of the 20 essays assigned: #8, #9, and #10. These essays were selected as they were those found mid-task, a point in time where the raters had already scored some texts and had warmed-up, but they were still not too tired to do what was asked of them (recording of thoughts verbalization and self-analysis of performance).

The subjects were asked to read a document prepared by the researchers before rating the essays. The guidelines specified on it asked the raters to rate the essays in the same order as they were numbered and to focus their self-analysis on three of the six categories extracted in Stage 1: cognitive processes through the rating process, task operationalization, and assessment tool, since these categories were the ones that could be proved more useful to meet the study objectives; also, they lent themselves to be modified in the short term if specialized training were given, and they could yield observable results in the assessment process implementation. In this stage of the process, those three categories were renamed as *declarative knowledge, procedural knowledge*, and *checklist / assessment instrument*, respectively. Each of them was explained as well as its purpose within the research project hereby described.

The guidelines also requested the subjects to share all their thoughts regarding variables such as possible perceived bias, troublesome indicators, assessment of construct, and transition in-between compositions, among others. Then, the researchers transcribed the information conveyed in these audios, analyzed it and classified it. The observations gathered were shared with PPEILE and other faculty members to provide stakeholders with a broader view of the analysis that was carried out to improve the evaluation process of a high-stake evaluation such as that of the entrance exam for PPEILE.

Subjects

The subject matter experts (SMEs) who took part in both stages of the study were two TEFL professors at UCR who are active collaborators of PELEx. Both SMEs have a BA in English, a master's degree in teaching English as a Foreign Language, and approximately 12 years of

experience in the field. SME 2's training in scoring writing tasks started five years ago. Her familiarity with CEFR and its language proficiency descriptors accounts for two years, approximately. SME 3 has four years of experience in grading writing tasks and has received extensive training in CEFR descriptors as this person is part of various initiatives that use these as guidelines for the creation and adaptation of assessment tools.

Results

The purpose of Stage 1 was to identify preliminary analysis categories in the rating process of the essay, which resulted in:

1. Emotional area
2. Input, prompt, and other resources
3. Work operationalization
4. Cognitive processes for the rating tasks
5. Observations about the assessment tools
6. Rater's profile

After Stage 2, the study focused on analysis categories #3, #4 and # 5 from the previous list. Each of them was renamed with a more precise theoretical term that elicited key elements which should guide the future design of rater training and calibration processes as shown in Table 4.

Table 4
Categories derived from the CTA process

Category Name (Phase 1)	Category Name (Phase 2)	ELEMENT
Work operationalization	**Procedural Knowledge**	• Scoring moment • Work area • Task length • Pre-task protocol • Task methodology • Task instruments and tools
Cognitive processes for the rating tasks	**Declarative Knowledge**	• Error recognition • Identification of omission error • Identification of incorrect use • Recognition of undesired elements • Recognition of examinee's weaknesses in written production • Recognition of correct elements in the task • Rater's methodology applied during task • Procedures used by raters that require further validation

Observations about the assessment instrument	**Assessment instrument**	• Comments about the components and indicators in the checklist • Suitability for the task • Recommendations to improve the instrument • Scoring inconsistencies

Limitations of the Study

Although CTA research highlights the importance of exploring the cognitive process of as many SMEs as possible since this would provide further certainty on the results obtained, the present study focused on very few subjects, as these were the only ones appointed for the task by the stakeholders and the few that were available at the moment this research took place.

Even though the pilot stage provided researchers with a departing point to determine the SMEs whose rating was closer based on the grade assigned to only one essay, very few pilot essays were part of the sample. In addition, the quality and quantity of the information provided by SMEs varies according to each individual and their context.

CTA has been commonly used in the planning stage of training programs since it provides valuable insight from expert sources regarding processes that are normally not verbalized. However, there seems to be very little research on its use as a validation tool for assessment. More studies on the value of this information elicitation technique should be carried out as source of data for subsequent rater calibration programs.

The verbalization of SMEs' thoughts and steps as they were rating the essays may have unforeseen consequences on their performance as this is an activity that they would not normally carry out. First, concern over the truthfulness of their statements may arise (*verticality*). Also, the attention given to their own rating could have brought about an over-analysis of their decisions and affected the normal pace of the process (*reactivity*). However, as Honko et al. (2023) and Barkaoui (2011) state, such concerns have already been debunked. In fact, the detailed attention given to their actions as the raters performed the task could have also helped them to be more aware of their judgment and, consequently, justify it with more sound reasoning.

Conclusions and Recommendations

As already mentioned, the outcomes expected from this CTA research described in this article will be seen in the long term. Several administrations of the test and their analysis must be carried out before observing real effects on a calibration of raters for written compositions, which may have an impact on their efficiency, consistency, and rating rigor, among others. Nonetheless, in the short term, the process has already produced some results that will allow further research and work to fulfill the main goals. This section includes comments drawn of the three chosen categories (declarative knowledge, procedural knowledge, and assessment instrument) in stage two and general suggestions derived from all the process.

Stage two allowed the elicitation of SMEs' mental processes, cognitive elements, reasoning, and operationalization of task as they performed the rating activity, which will eventually be taken as particular aspects of a training workshop for new judges in future evaluations. Three categories were eventually focused on: declarative knowledge, procedural knowledge, and assessment instrument.

For the first category, researchers determined that to work on rater calibration, raters must be first trained on terminology and theoretical criteria related to the writing skills and its evaluation. To do this effectively, stakeholders and other writing experts must work on a standardization of key concepts that every rater must possess along with underdeveloped criteria that would not be pointed out by SMEs in this and future works.

In terms of procedural knowledge, a significant takeaway obtained from the analysis of experts' thoughts is a clear need to set guidelines for future raters in terms of timespan and workspace choices, management of external distractions, and transitional activities in-between grading compositions. These guidelines should also include suggested gadgets and tools to be used. Researchers also identified some controversial or debatable procedures which require further analysis in order to determine whether they should be recommended or warned about in the guidelines.

Finally, regarding the assessment instrument, the analysis uncovered the lack of indicators to rate the concluding paragraph, which prevented the judges from rating the essays from a more comprehensive scope. SMEs' comments also revealed some inconsistent understanding of a couple of indicators and terms used within them; such misconstructions must be addressed in future training endeavors. There is also an urgent action, by test designers, to conduct an analysis of compliance of the instrument with respect to the assessment construct.

As for general takeaways from the CTA process explained in this article, the following is a list of the most significant findings:

- SMEs' professional background must be a selection criterion.
- Ongoing modifications to the checklist, after each administration of the examination, are needed to ensure its adherence to the construct and to be more user-friendly.
- Prompt design is paramount to meet task objectives successfully.
- The affective component is continuously addressed by all raters throughout the process, which brings out the noteworthiness of the inclusion of this category in future CTA research.
- To identify and work with more categories, experts on other inter-related fields must become involved.
- During this CTA study, raters' performance was divergent in varied areas such as operationalization, reasoning, and scoring, among others. This is solid evidence of the need to construct protocols and training workshops to calibrate raters for future assessments in written composition.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (Eds.), (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing 28*(1), 51-75. https://doi.org/10.1177/0265532210376379

Burgos, M., López-Martín, M., Aguayo-Arriagada, C., & Albanese, V. (2022). Análisis cognitivo de tareas de comparación de probabilidades por futuro profesorado de Educación Primaria. *Uniciencia, 36*(1), 588-611. https://dx.doi.org/10.15359/ru.36-1.38

Cannon-Bowers, J., Bowers, C., Stout, R., Ricci, & K., Hildabrand, A. (2013). Using cognitive task analysis to develop simulation-based training for medical tasks. *Military Medicine, 178*(10), 15-21. https://doi.org/10.7205/MILMED-D-13-00211

Clark, R., Feldon, D, Van Merriënboer, J., Yates, K, & Early, S. (2008). Cognitive task analysis. In Spector, J, Merril, M.D., Van Merriemboer, J., Driscoll, M. (Eds.), *Handbook of Research on Educational Communications and Technology* (3rd ed., pp.577-593). Routledge.

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Press Syndicate of the University of Cambridge.

Crandall, B., Klein, G. A., & Hoffman, R. R. (2006). *Working minds: A practitioner's guide to cognitive task analysis*. MIT Press.

Gallagher, J.P. (1979). Cognitive/information processing psychology and instruction: Reviewing recent theory and practice. *Instructional Science*, 8, 393-414. http://www.jstor.org/stable/23368243

Hoffman, R., Crandall, B., & Shadbolt, N. (1998). Use of the critical decision method to elicit expert knowledge: A case study in the methodology of cognitive task analysis. *Human Factors*, *40*(2), 254-276. https://doi.org/10.1518/001872098779480442

Hoffman, R., & Militello, L. (2008). *Perspectives on Cognitive Task Analysis: Historical Origins and Modern Communities of Practice*. Psychology Press.

Honko, M., Neittaanmäki, R., Jarvis, S., & Huhta, A. (2023). Beyond literacy and competency – The effects of raters' perceived uncertainty on assessment of writing. *Assessing Writing, 57*. https://doi.org/10.1016/j.asw.2023.100768

Klein, G. A., Calderwood, R., & MacGregor, D. (1989). Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, & Cybernetics, 19*(3), 462–472. https://doi.org/10.1109/21.31053

Klein, G., & Militello, L. (2001). Some guidelines for conducting a cognitive task analysis. In E. Salas (Ed.), *Advances in human performance and cognitive engineering research* (pp. 163–199). Elsevier Science/JAI Press. https://doi.org/10.1016/S1479-3601(01)01006-2

Knisely, B. M., Joyner, J. S., & Vaughn-Cooke, M. (2021). Cognitive task

analysis and workload classification. *MethodsX*, *8*(101235). https://doi.org/10.1016/j.mex.2021.101235

Lyness, S.A., Peterson, K., & Yates, K. (2021). Low Inter-Rater Reliability of a High Stakes Performance Assessment of Teacher Candidates. *Education Sciences,*11, 648. https://doi.org/10.3390/educsci11100648

McManus, K. (Ed.). (2024). *Usage in Second Language Acquisition: Critical Reflections and Future Directions* (1st ed.). Routledge.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Militello, L., & Hutton, R. (1998). Applied cognitive task analysis (ACTA): a practitioner's toolkit for understanding cognitive task demands. *Ergonomics*, 41(11), pp. 1618-1641. https://doi.org/10.1080/001401398186108

Otálora , Y. (2019). El análisis cognitivo de tareas como estrategia metodológica para comprender y explicar la cognición humana. *Universitas Psychologica, 18*(3), 1-12. https://doi.org/10.11144/Javeriana.upsy18-3.acte

Schraagen, J. M., Chipman, S. F., & Shalin, V. L. (Eds.). (2000). *Cognitive task analysis*. Psychology Press.

Snow, E. & Katz, I. (2009). Using Cognitive Interviews to Validate an Interpretive Argument for the ETS iSkills™ Assessment. *Communications in Information Literacy 3*(2), 99-128. http://dx.doi.org/10.15760/comminfolit.2010.3.2.75

Solís Hernández, M. (2009). Graduates' Degree of Satisfaction with the MA Program in Teaching English as Foreign Language at the University of Costa Rica. *Revista de Lenguas Modernas*, (10). https://revistas.ucr.ac.cr/index.php/rlm/article/view/8900

Tofel-Grehl, C. & Feldon, D. (2013). Cognitive Task Analysis-Based Training: A Meta-Analysis of Studies. *Journal of Cognitive Engineering and Decision Making*. 7. 293-304. https://doi.org/10.1177/1555343412474821

Watkins, R.; & Visser, Y. (2009). Cognitive Task Analysis. In Biech, E. (Ed.), *The 2009 Pfeiffer Annual: Talent Management*. Jossey-Bass/Pfeiffer. https://www.learndev.org/People/YusraVisser/2008_VisserWatkins_CogTaskAnalysisPfeiffer.pdf