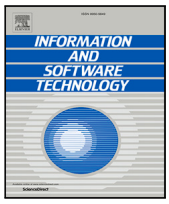
Information and Software Technology 178 (2025) 107624

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof



Evaluating the understandability and user acceptance of Attack-Defense Trees: Original experiment and replication ☆

Giovanna Broccia [a,*], Maurice H. ter Beek [a], Alberto Lluch Lafuente [b], Paola Spoletini [c], Alessandro Fantechi [d,a], Alessio Ferrari [a,*]

[a] CNR–ISTI, Pisa, Italy
[b] DTU Compute, Technical University of Denmark, Kongens Lyngby, Denmark
[c] Kennesaw State University, GA, USA
[d] DINFO, University of Florence, Florence, Italy

ARTICLE INFO

Keywords:
Security requirements
Attack-Defense Trees
Understandability evaluation
Users acceptance
Empirical user study
Internal replication
Method Evaluation Model

ABSTRACT

Context: Attack-Defense Trees (ADTs) are a graphical notation used to model and evaluate security requirements. ADTs are popular because they facilitate communication among different stakeholders involved in system security evaluation and are formal enough to be verified using methods like model checking. The understandability and user-friendliness of ADTs are claimed as key factors in their success, but these aspects, along with user acceptance, have not been evaluated empirically.
Objectives: This paper presents an experiment with 25 subjects designed to assess the understandability and user acceptance of the ADT notation, along with an internal replication involving 49 subjects.
Methods: The experiments adapt the Method Evaluation Model (MEM) to examine understandability variables (i.e., effectiveness and efficiency in using ADTs) and user acceptance variables (i.e., ease of use, usefulness, and intention to use). The MEM is also used to evaluate the relationships between these dimensions. In addition, a comparative analysis of the results of the two experiments is carried out.
Results: With some minor differences, the outcomes of the two experiments are aligned. The results demonstrate that ADTs are well understood by participants, with values of understandability variables significantly above established thresholds. They are also highly appreciated, particularly for their ease of use. The results also show that users who are more effective in using the notation tend to evaluate it better in terms of usefulness.
Conclusion: These studies provide empirical evidence supporting both the understandability and perceived acceptance of ADTs, thus encouraging further adoption of the notation in industrial contexts, and development of supporting tools.

1. Introduction

Defining security requirements involves representing and analysing potential threats and mitigation strategies to establish a security policy [1]. Several notations have been developed in requirements engineering (RE) to model and analyse security requirements. These include extensions of well-known notations, such as Secure i* [2] and Secure UML [3], as well as comprehensive notations with analytical capabilities, like the Socio-Technical Security Modelling Language (STS-ML) [4] and the Restricted Misuse Case Modeling (RMCM) approach [5].

Among these various approaches, Attack-Defense Trees (ADTs) provide a graphical notation for modelling and assessing the security requirements of systems or assets. ADTs provide a graphical, tree-based representation of the possible actions an attacker might take to compromise a system and the corresponding defensive measures [6].

The purposes of ADTs are manifold: they offer a comprehensive threat modelling methodology, enable users to graphically identify potential threats and corresponding security strategies, and allow for

☆ Research supported by the Italian MUR–PRIN, Italy 2020TL3X8X project T-LADIES (Typeful Language Adaptation for Dynamic, Interacting and Evolving Systems); by Innovation Fund Denmark and the Digital Research Centre Denmark, through the bridge project "SIOT – Secure Internet of Things – Risk analysis in design and operation"; by Industriens Fond, Denmark through the project "Sb3D: Security-by-Design in Digital Denmark"; and by the EU Horizon Europe project CODECS GA 101060179. The authors would like to thank all the participants of the study.

* Corresponding author.
E-mail addresses: giovanna.broccia@isti.cnr.it (G. Broccia), maurice.terbeek@isti.cnr.it (M.H. ter Beek), albl@dtu.dk (A. Lluch Lafuente), pspoleti@kennesaw.edu (P. Spoletini), alessandro.fantechi@unifi.it (A. Fantechi), alessio.ferrari@isti.cnr.it (A. Ferrari).

https://doi.org/10.1016/j.infsof.2024.107624
Received 24 June 2024; Received in revised form 23 September 2024; Accepted 31 October 2024
Available online 9 November 2024

the quantitative assessment of a system's security level using formal verification techniques such as model checking. Furthermore, ADTs facilitate communication among stakeholders from diverse backgrounds, such as security experts and system engineers.

ADTs have been touted as one of the most popular graphical models for system security analysis [7]. They are acclaimed for their user-friendliness, even for novices [8], and their easily understandable, human-readable notation [9]. Despite these claims, no empirical studies have been conducted to verify their validity, even though this research direction has been recognised as promising and would be valuable for evaluating the effectiveness of ADTs [7,10,11].

The understandability of graphical notations is a well-studied area in the literature [12,13], particularly in the context of security modelling and assessment [14–16]. These studies are particularly beneficial given the central role of humans in system security, encompassing both potential insider threats and human errors that can render systems vulnerable [9]. Effective security modelling notations must be easily understood by users to ensure accurate implementation and response strategies.

User acceptance of notations or methods is equally critical, as the literature suggests that perceived acceptance can predict actual usage [17–19].

In this paper, we present the first evaluation that aims at investigating the quality of the ADT notation, both in terms of understandability and in terms of user acceptance, through two empirical studies. The first experiment was conducted entirely online, providing broad accessibility and diverse participant engagement. The second experiment was an internal differentiated replication conducted in person with different subjects, offering a controlled environment and in-depth interaction. In the field of empirical software engineering, replications are crucial for enhancing the validity of results [20,21]. The proposed replicated experiment enables a deeper evaluation of ADTs, thus contributing to the field of security RE.

In our analysis, we tailored the Method Evaluation Model (MEM) [18,22] to evaluate the understandability and user acceptance of the ADT notation. Understandability of ADTs was measured through a test in which participants were required to perform a set of tasks related to syntax, semantics, and usage of ADTs. User acceptance was measured through a questionnaire, evaluating perceived ease of use, perceived usefulness, and intention to use.

Our results show that: (1) ADTs are well understood by participants; (2) ADTs are perceived as easy to use and useful, with participants expressing a strong intention to use them; (3) there is a significant relationship between perceived usefulness and the intention to use ADTs; and (4) there are no significant relationships between various performance-based measures of understandability (effectiveness and efficiency) and perception-based variables (ease of use, usefulness, intention to use), except in the following case: those who *apply* the method better in practice also consider it more useful.

The present paper builds upon and extends the work in [23]. The previous study presented the first experiment with 25 participants and was conducted entirely on an online platform without any interaction with the examiners. This work was exploratory in nature and contributed with a first indication of the understandability and acceptance of ADTs. The present paper, and the second experiment in particular, is confirmatory in nature, and offers the following extensions:

- **Internal differentiated replication.** The study introduces a differentiated replication of the original study presented in [23]. The replication involves 49 participants, and was conducted in person (Sections 5.2 and 6.2). While the experimental phases, variables, and data analysis remained consistent with the original study, adjustments were made to the materials to accommodate the in-person format. This replication offers more robust data and comparisons, enhancing the reliability and generalisability of the findings.
- **Comparative analysis of the experiments.** A detailed comparative analysis of the results from the two experiments is presented (Section 6.3), offering new insights into the consistency of the findings across different settings and participant groups.
- **Comprehensive discussion of the results.** We provide a more extensive discussion of the combined results from both studies (Section 6.4), shedding light on broader trends and possible implications.
- **Expanded review of related work.** We offer a more comprehensive and structured review of the literature (Section 3), covering key areas such as security modelling notations, empirical evaluations of graphical security methods, and frameworks used to assess understandability and user acceptance. This expanded review deepens the contextualisation of ADTs within the broader security modelling landscape and strengthens the foundation for the empirical study.

The contributions of this study are as follows:

- **Consolidated evidence regarding ADT understandability and acceptance.** The results of the two experiments offer evidence supporting the high level of understandability and user acceptance of ADTs, thus providing data to confirm previous anecdotal claims. This contribution is relevant to encourage further adoption of the notation in industrial settings, and further development of supporting tools.
- **Contribution to the field of conceptual modelling.** By showing that ADTs are understandable and accepted by users, the study offers valuable insights into the practical application of conceptual RE models, such as ADTs. Although conceptual RE models, especially formal ones, are not frequently adopted in practice [24], our results suggest that this phenomenon may not be due to inherent weaknesses of the notations, but rather to other contextual factors, such as prejudice, or lack of integration into existing processes, as observed for other formal tools [25]. On the other hand, our results foster the study of the reasons for the success of ADTs, which can offer guidance for the development of other graphical notations.
- **Contribution to the field of empirical RE.** This paper presents a replication research design (Section 4) and package [26]. By offering a structured approach to replication, and exhaustive experimental material, the paper gives other researchers the possibility to perform external replications, which may extend the scope of validity of our conclusions.

The remainder of this paper is structured as follows. After a presentation of the background information and the discussion of related work in Sections 2 and 3, respectively, Section 4 outlines the experimental design, while Section 5 details the individual experiments. Section 6 presents the results of both experiments, including the comparative analysis and discussion. Section 7 addresses the threats to validity, and finally, Section 8 presents the conclusions and suggests directions for future work.

Replication Package. Our replication package is publicly available [26].

2. Background

In this section, we provide the relevant background on Attack Defense Trees and the Method Evaluation Model.

2.1. Attack-defense trees

The assessment of system security through graphical tree-based structures originated around 1960 with fault tree analysis [27], and gradually spread with the usage of similar structures such as attack trees [28,29]. To manage the dynamic nature of system security,

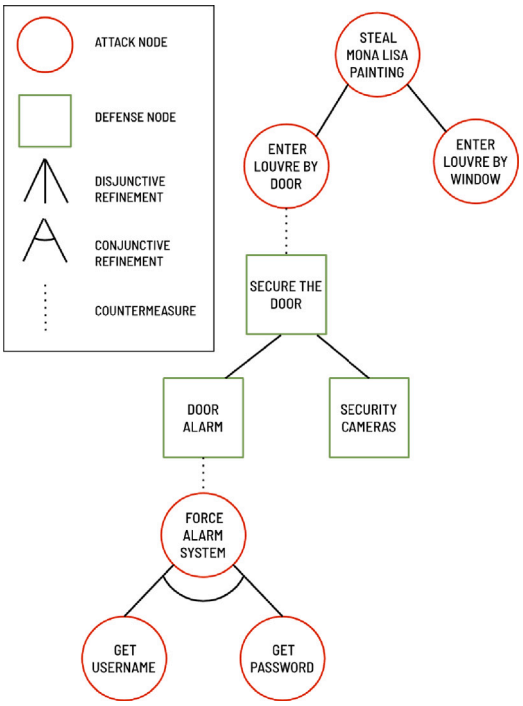


Fig. 1. ADT for theft of Mona Lisa.

Attack-Defense Trees (ADTs) [6] were introduced, extending attack trees with defense strategies and quantitative risk assessment [30,31]. ADTs model attack-defense scenarios, namely 2-player games between a proponent and an opponent.

The legend of Fig. 1 reports the basic graphical elements of ADTs, while the rest of the figure presents an example scenario that will be described later in this paragraph. Formally, ADTs are rooted trees with labelled nodes of two opposite types: attack nodes and defense nodes, representing the goals of the attacker and of the defender, respectively. The root can be of either type: if the root is an attack node, the proponent is an attacker; conversely, if the root is a defense node, the proponent is a defender. The main goal can be refined into sub-goals, described by its child nodes of the same type. The refinement can be either conjunctive (i.e., all sub-goals must be achieved to achieve the parent goal) or disjunctive (i.e., at least one of the sub-goals must be achieved to reach the parent goal). A node with no children of the same type is called a non-refined node, and it represents a basic/atomic action. Each node may have one child of the opposite type, representing a countermeasure to its (sub-)goal. Essentially, an attack node may have a number of children that refines the attack and a single defense node that fends it off. Conversely, a defense node may have a number of children which refines the defense, and a single attack node that counterattacks it.

To demonstrate the features of ADTs, we present a simple fictitious scenario describing the theft of the Mona Lisa painting (cf. Fig. 1). To steal the painting (see root of the ADT), the attacker can carry out two kinds of attacks: enter the Louvre museum by the door or by the window (attack nodes "Enter Louvre by Door" and "Enter Louvre by Window", forming a disjunctive refinement of the root node). Fig. 1 shows in detail only the door branch—further attacks and defenses could easily be added. The defender can secure the door of the museum (defense node "Secure the Door" used as a countermeasure for the attack node "Enter Louvre by Door") using a "Door Alarm" or "Security Cameras" (defense nodes, disjunctive refinement of the "Secure the Door" node). The attacker can then perform a counterattack by forcing the alarm system (attack node "Force Alarm System" used as a countermeasure for the defense node "Door Alarm"). To do so, the attacker needs to get both the username and the password (conjunctive refinement of the "Force Alarm System" node).

Due to their theoretical underpinning, ADTs enable formal reasoning, typically supported by effective software tools, on quantitative risk assessment (e.g., to determine where defensive resources are best spent). Academic tools like ADTool [30], SPTool [32], ATTop [33], and RisQFLan [31], as well as commercial tools such as AttackTree,[1] Risk*Tree*,[2] and Secur*IT*ree [34].[3] They are used in public and private sectors (e.g., by aerospace, defense and intelligence organisations, but also by health care providers, critical infrastructure companies, and financial organisations).

Evaluation of ADTs has so far considered issues like the consistency between an ADT and the system and the impact of repeated labels on results [35,36]. As far as we know, there is no work in the literature that has focused on the assessment of the comprehensibility of ADTs (neither of attack trees). Albeit their comprehensibility is usually assessed as a factor of success [7–9].

2.2. Method evaluation model

The Method Evaluation Model (MEM) [18,22] is an acceptance model used to evaluate new information technologies , extending the Technology Acceptance Model (TAM) [17] by incorporating performance measures. Essentially, MEM integrates the concepts of *perceived success* and *actual success* to predict the real usage of a method.

Perceived success involves assessing users' perceptions of the method's effectiveness and it is measured through a combination of three variables: the *perceived ease of use* (PEOU), which measures how easy the technology is perceived to be, the *perceived usefulness* (PU), which measures how useful the technology is perceived to be, and the *intention to use* (ITU), which measures the extent to which users intend to use the technology in the future. The combination of these three variables indicates the overall *acceptance* of the method.

Actual success entails evaluating the extent to which users effectively utilise the method and it is measured through performance-based variables consisting of efficiency and effectiveness, which measure the effort required to use the technology and how well the technology has been used to reach the goals, respectively. These variables must be tailored by delineating specific objectives for the method under analysis.

MEM has been applied in the fields of RE [37,38] and language comprehension [39]. In [38], the performance-based variables (effectiveness and efficiency) were adapted to measure the ability of identifying security threats. Conversely, in [37,39], effectiveness and efficiency were tailored to measure the *understandability* of requirement models and language constructs, respectively. In practice, the performance-based variables are understandability effectiveness and understandability efficiency, computed based on the results obtained by sample subjects in problem-solving tasks. This paper adopts this latter approach and further decomposes the variables into fine-grained dimensions (cf. Section 4.1). In line with MEM, we evaluate if these variables are related to perception-based variables.

3. Related work

This section reviews related work on security notations, the empirical evaluation of graphical security methods, and understandability models.

[1] https://www.isograph.com/software/attacktree/.
[2] https://risktree.2t-security.co.uk/.
[3] https://www.amenaza.com/.

3.1. Security modelling and analysis notations

Several notations have been proposed in RE to model and analyse security requirements [40–43]. Most of these notations are extensions of existing RE notations, like Secure i* [2], an extension of KAOS [44], Secure UML [3], Misuse cases [45], and Secure Tropos [46]. Other authors propose entirely novel methods, as, e.g., CORAS [47]. In the following, we describe the main contributions to the field of security requirements modelling.

Secure i* [2] extend i*, a well-known goal-oriented RE method focused on social modelling, by putting emphasis on potential malicious players and associated goals, and enabling the analysis of weak social links that can be exploited by attackers and system abusers. A preliminary evaluation of the language was performed by the authors on example cases. Still focused on goal modelling, the approach by Salehie et al. [44] enhances KAOS goal-oriented models by introducing asset models and threats models. The three models are used as input to build a causal network to analyse system security, and enable the identification of countermeasures to security threats. The approach is evaluated through simulation on three different example scenarios. Secure UML [3] focuses on the security of distributed systems, and the enforcing of access control. To this end, it defines a vocabulary for annotating UML class diagrams with information relevant to access control. The approach is showcased on an example case. Misuse cases [45] are an extension of UML Use Cases that introduce possible undesired actions that can be performed by attackers, and enable the definition of associated mitigations. The approach provides both graphical and textual specifications, and has been employed in different European projects. The Restricted Use Case Modeling method (RUCM) [5] is a use case-driven modelling method that uses misuse case diagrams [45] to support the specification of security and privacy requirements of multi-device software ecosystems in a structured and analysable form. The approach is evaluated on an industrial healthcare project, and feedback was acquired through a questionnaire and interviews with four engineers. Secure Tropos [46] extends Tropos, a development methodology using the i* modelling framework, by introducing additional security-focused concepts, such as security constraints, entities, trust relationships between actors, and others. The approach is demonstrated on an real-world case from the healthcare domain. A further extension of Tropos, named STS-ml [4] provides an actor- and goal-oriented security requirements modelling language, able to capture system security needs and requirements at the organisational level and reason about corporate assets, social dependencies, and trust properties. The requirements models of STS-ml incorporate formal semantics, facilitating automated reasoning to identify potential conflicts among security requirements and between security requirements and actors' business policies. The proposal has been implemented in the STS-Tool, and the method has been quantitatively evaluated on a real-world case for (a) its capability of identifying non-trivial conflicts, and (b) its scalability when applied to large models. Finally, CORAS [47] is a complete modelling and risk analysis method composed of eight steps that not only enables reasoning on security aspects, but also consider any type of risk and vulnerability. The method distinguishes itself from others also because an extensive manual with several examples has been published to facilitate its use [48].

Overall, these works show that problem of modelling security requirements has been addressed by multiple perspectives, e.g., social [2,4], use case [5,45], access [3], asset [44], or process [47]. Although forms of empirical evaluation exist for some of the methods [4,5], the solution proposals did not include controlled experiments with human subjects to evaluate quality aspects of the modelling languages, as, e.g., their understandability. However, other authors have considered the available notations and performed such experiments. In the following, we summarise the main contributions in this sense.

3.2. Empirical evaluation of graphical security modelling methods

A substantial number of empirical studies analysing security requirements modelling and representation methods was carried out by the team composed of Massacci, Paci, Labunets et al. [15,16,38,49,50]. In the first study [49], the authors considered different academic security risk assessment methods, including CORAS, Secure Tropos, Secure i* and Problem Frames [51]—an RE approach not specifically tailored for security requirements, but previously employed by Haley et al. [52] to support security analysis. They involved MSc students in computer science and professionals in IT Audit for Information Systems, who had no previous knowledge of the methods, and, among other quality factors, also compared the usability of the different tools. Based on the evidence, they conclude that CORAS is substantially more usable than the other solutions. In the following studies [38,50], they performed controlled experiments to compare the CORAS visual notation and framework with a textual method used for air traffic control security assessment, named SecRAM [53]. Similar to our study, they adapted MEM and TAM for evaluating PEOU, PU, and ITU of both methods, and concluded that the visual method is better perceived by users across the different variables. However, concerning their effectiveness in identifying security threats, the two approaches can be considered equivalent. This finding was confirmed by a later study [15]. However, the same study showed that also user perception variables were equivalent for both methods. CORAS was also compared with another textual method, named SREP [54]. The study concluded that the visual method is more effective for identifying threats than the textual one and is more appreciated by users, whereas the textual method is slightly more effective for eliciting security requirements.

Volden-Freberg et al. [55] conducted an empirical study to evaluate graphical versus textual risk annotations in threat models represented through UML sequence diagrams. The aim of the study was to compare the comprehensibility of the two methods in analysing security threat problems, as well as to assess the efficiency of these two annotation types by measuring the average time each group spent per task. Their findings indicate that while graphical and textual annotations show comparable comprehensibility, the graphical method proved to be more efficient than the textual method.

In [10], the results of an empirical evaluation conducted to determine the effectiveness of two attack modelling techniques, an adapted attack graph method based on [56] and the fault tree standard – a notation not strictly focused on security requirements – are reported. Similar to our study, the objective was to test the participants' ability to recall, comprehend, and apply these techniques. The results indicate that the attack graph method is more effective than the fault tree method, suggesting that specialised modelling solutions targeting security requirements are preferable to general-purpose ones. Furthermore, participants with a computer science background performed better than those without experience when using both methods.

In [57], a graphical approach to facilitate communication and understanding among different classes of users during a risk analysis brainstorming session was proposed. The development and the guidelines for the use of such a graphical language were based on a combination of empirical investigations and experiences gathered from utilising the approach in large-scale industrial field trials by both professionals and students.

Previous works on the evaluation of security modelling methods mainly focus on the comparison between textual and graphical approaches, and assess their effectiveness in identifying security threats, as well as their acceptance by users. These studies mainly focus on the CORAS graphical approach. A more limited set of studies, i.e., [10,55], consider also the understandability of existing notations, and do not account for acceptance. We are not aware of studies that evaluate these quality aspects for the majority of the notations mentioned in Section 3.1. Our work differs from these previous contributions in that (1) considers both understandability and acceptance and (2) focuses on the ADT notation, which is widely used in industry.

3.3. Understandability models

Understandability is a critical factor in the effective use of modelling notations and conceptual systems across software engineering, information systems, and various technical domains. Various models and frameworks have been developed to assess both the objective and the perceived understandability of these notations. Objective understandability focuses on measurable outcomes such as task effectiveness and efficiency, while perceived understandability relies on participants' subjective assessments, often emphasising perceived ease of use, usefulness, intention to use, and cognitive load [58].

These models help researchers evaluate the understandability of models by examining factors such as cognitive effort, ease of use, and clarity of representation. Each model offers a distinct approach, with some emphasising cognitive aspects and others focusing on user perception, particularly in terms of usability, usefulness, and ease of use.

Among the models emphasising cognitive aspects one of the most widely adopted ones is the Cognitive Dimensions of Notations (CDN) framework, introduced by Green and Petre [59]. This framework provides a set of cognitive dimensions that help identify usability trade-offs in a notation system, focusing on how users interact with and understand the structure of a model. CDN helps identify how a notation's structure affects users' cognitive efforts in understanding it.

Similarly, the Cognitive Load Theory (CLT) [60] has been applied to evaluate the understandability of models by measuring the mental effort required for users to understand them. CLT categorises cognitive load into intrinsic load, which is related to the inherent complexity of the material; extraneous load, which refers to the cognitive effort imposed by the way information is presented; and germane load, which concerns the effort invested in constructing and understanding new knowledge. In modelling notations, reducing extraneous load – by simplifying the presentation and organisation of information – can significantly improve understandability.

On the other hand, models focusing on perceived understandability are commonly known as acceptance models. Technology acceptance research has been a mature field for over two decades (cf. [61] for an overview on popular models and theories). Several theoretical models, primarily developed from psychological and sociological theories, explain technology acceptance and use [62]. One of the earliest such models, the Theory of Reasoned Action (TRA), introduced by Fishbein and Ajzen in 1975 [63], explains how an individual's behavioural intentions are influenced by two key factors: attitude towards the behaviour, which refers to the individual's positive or negative feelings about performing the behaviour, and subjective norms, which involve the social pressure or influence from others regarding whether the individual should perform the behaviour.

Later developments in social cognitive theory provided further insights into understandability. In 1986, Bandura introduced Social Cognitive Theory (SCT) [64], which theorises that learning occurs in a social context with a dynamic and reciprocal interaction between environmental factors and behaviours. To improve upon some of the drawbacks of TRA, Ajzen conceived the Theory of Planned Behaviour (TPB) in 1991 [65], adding the determinant of perceived behavioural control, which refers to the perception of how easy or difficult it is to perform a given behaviour.

TRA served as the foundation for TAM [17], which introduced perceived ease of use as a critical factor closely related to understandability. TAM has been widely used to assess how easily users can understand and interact with new systems, including conceptual models.

In 2003, Venkatesh et al. [19] developed the Unified Theory of Acceptance and Use of Technology (UTAUT), which integrated components from eight major technology acceptance models and theories, including TRA, TAM, TPB, and SCT. UTAUT identified four primary determinants of usage and intention: performance expectancy, effort expectancy, social influence, and facilitating conditions. These are moderated by gender, age, experience, and voluntariness of use, providing a comprehensive framework for understanding how perceived understandability impacts the adoption of models and technologies. UTAUT has been shown to perform better in certain contexts, particularly in organisational settings where factors like social influence and facilitating conditions play a key role in predicting technology adoption [19]. Additionally, it is more effective in mandatory usage scenarios, such as workplace environments, where performance expectancy and effort expectancy are influenced by external pressures [62]. The model also excels in longitudinal studies, as it accounts for variables like experience and voluntariness of use, which evolve over time, enhancing its predictive accuracy [66]. Finally UTAUT demonstrates better performance in the context of mobile Internet users in consumer settings, providing the best explanation power for the intention to use and actual use of mobile Internet, outperforming other models [67].

MEM, proposed by Moody in 2001 [22], extends TAM and focuses on the effectiveness, efficiency, and acceptance of using modelling methods. This dual evaluation helps capture both the objective and subjective aspects of understandability, making it particularly useful for empirical studies on model evaluation.

3.4. Implications of previous work

The results of previous research have informed the design of the current study as follows. Previous work shows that many security modelling notations have been proposed, with only few evaluated empirically, mainly comparing textual and graphical approaches. These studies conclude that graphical notations, such as CORAS, are more accepted and more effective than textual ones. This justifies our focus on a graphical notation such as ADT. As noted in Section 2, this is widely used in practice and claimed to be easy to use, but no sound evaluation of its understandability and acceptance has been carried out. Among the different understandability models available, we adopted MEM since it has already been applied in the field of RE [37,38], which consolidates its suitability for evaluating modelling methods within this domain.

4. Experimental design

Our experimental design follows the guidelines by Wohlin et al. for software engineering experiments [68]. Our overarching goal is the following:

> **Goal:** Assessing the understandability and acceptance of ADTs by novice users with little to no prior knowledge of the notation, and exploring whether a relationship exists between the degree of acceptance of the notation and its understandability.

To address this goal, we perform two experiments, one referred to as original experiment and the other one as replication. The original experiment was conducted entirely online, without any direct interaction with participants. While this approach could yield positive outcomes, such as reduced subject bias, it might also introduce negative outcomes due to potential technological problems. Consequently, we replicated the experiment in person using a written format to mitigate these issues.

Based on the goal we derive the following research questions (RQs):

RQ1 *How well do novice users with no or minimal prior knowledge of the ADT notation understand ADTs?* This RQ aims at understanding the level of effectiveness and efficiency with which users, who lack a specific background in ADT notation, comprehend the notation

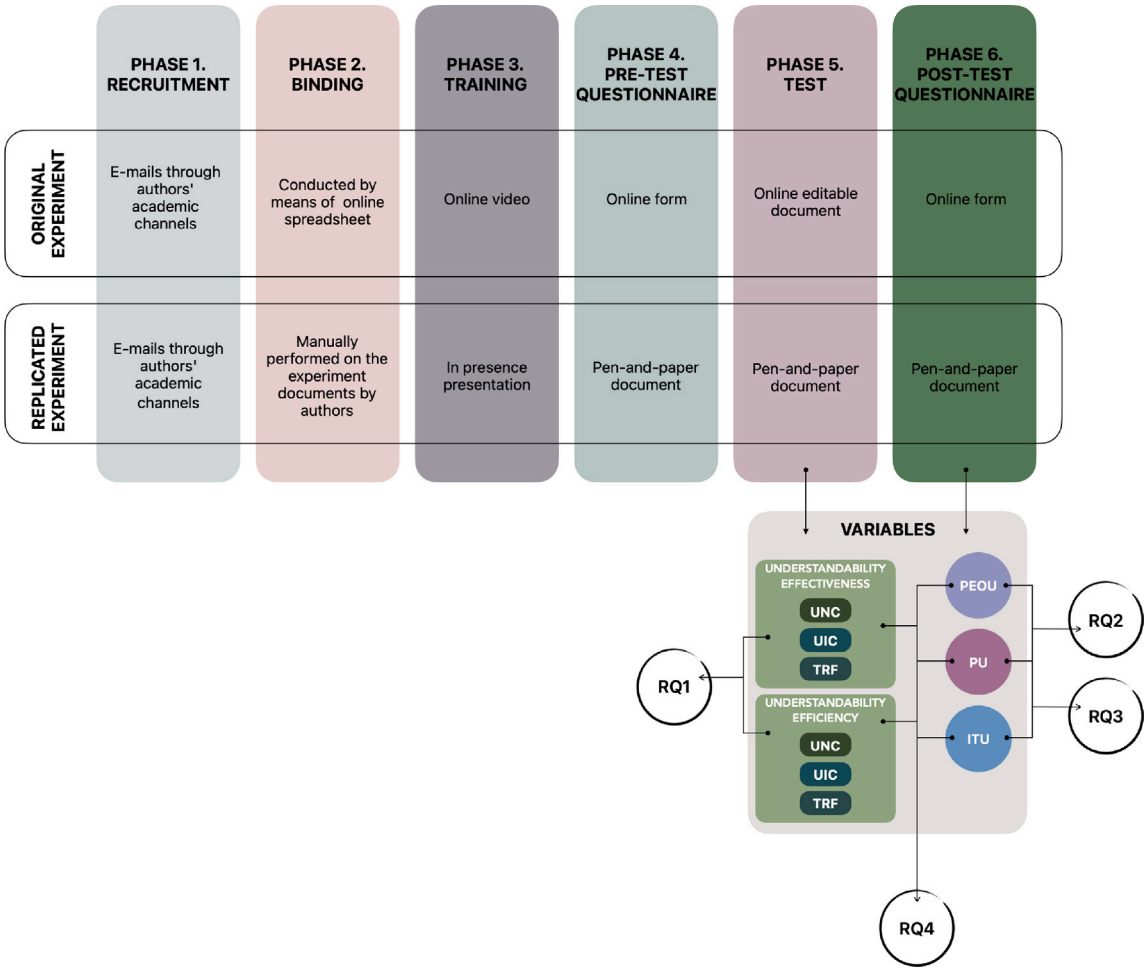


Fig. 2. Diagram illustrating the phases of the experiment in both the original and replicated experiment, the measured variables, and how these variables address the research questions.

RQ2 *What is the degree of acceptance of ADTs by novice users with no or minimal prior knowledge of the notation?* This RQ aims to evaluate how novice users perceive ADTs in terms of ease of use and usefulness, and the extent to which they would intend to adopt ADTs in the future. The focus is on assessing acceptance from a broader set of users, acknowledging ADTs' potential to facilitate communication among stakeholders with diverse backgrounds and skill sets.

RQ3 *What is the relationship between ease of use/usefulness of the notation and the intention to use it in the future?* Differently from RQ2, which focuses on each perception-based variable independently, this RQ aims at checking whether there is a relationship among the variables, and in particular, if ease of use and usefulness are related to intention to use.

RQ4 *What is the relationship between the ADT understandability and the users' perception of ADTs' ease of use and usefulness?* With this RQ, we check whether users who perform best in understanding the notation also tend to evaluate the ADTs as easier and more useful.

Both the replicated and the original experiment retain the same research questions; however, in this paper, RQ4 encompasses both RQ4 and RQ5 from the original contribution in [23].

We evaluate understandability in terms of effectiveness and efficiency through a test composed of a set of problem-solving tasks related to syntax, semantics, and usage of ADTs. Effectiveness measures how participants score on the test, while efficiency measures the effectiveness with respect to the time required to perform the test. Acceptance is evaluated in terms of perceived ease of use, perceived usefulness, and intention to use through a questionnaire adapted from MEM. These variables are described in detail in Section 4.1. Based on the RQs and the variables, we define a set of NULL hypotheses, described in Section 4.2.

Fig. 2 shows a diagram which illustrates our experimental design, namely the experimental phases, how they are instantiated in each experiment, the variables measured, and how those variables are associated with the RQs. Details about the different phases are reported in Section 4.3. The process of data analysis, oriented to statistically test the formulated hypotheses, is reported in Section 4.4.

The details on participants and resources of the individual experiments are presented in Section 5. The results and the analysis of the study validity are presented in Sections 6 and 7, respectively.

4.1. Variables and materials

The constructs, variables, materials used to assess them, and the methods of measurement are summarised in Table 1. Fig. 3 shows the adapted MEM with all the variables measured in the study and their relationship.

Acceptance. The evaluation of users' acceptance is based on the MEM model presented in Section 2.2. In particular, we evaluate acceptance using three perception-based variables: perceived ease of use (PEOU), perceived usefulness (PU), and intention to use (ITU).

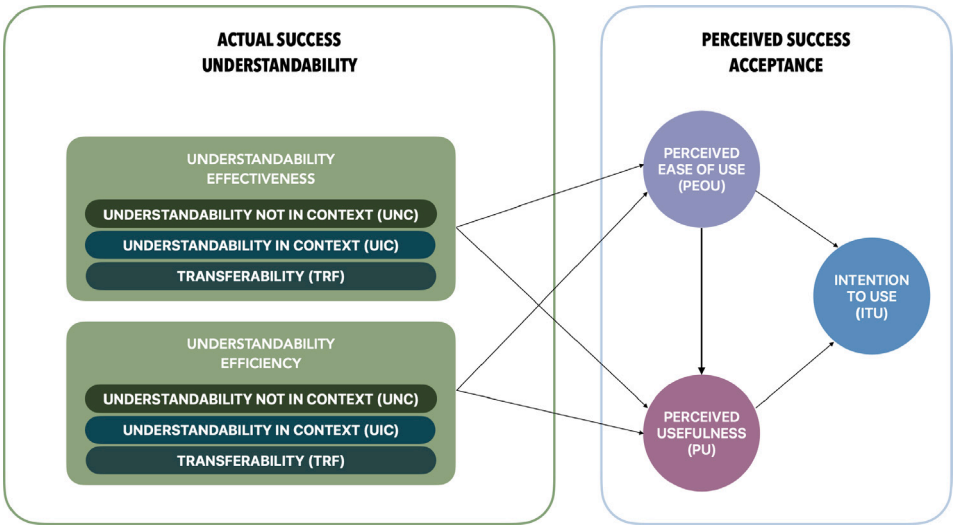


Fig. 3. Adapted MEM.

Table 1
Summary of variables.

Construct	Variable	Measure	Material
Understandability	Understandability not in context Effectiveness (UNC effv.)	$\sum$ *Correct Answers/Number of T/F Statements* (24)	6 ADT fragments with 4 T/F statements each
	Understandability not in context Efficiency (UNC effc.)	*UNC Effectiveness/Time*	
	Understandability in context Effectiveness (UIC effv.)	$\sum$ *Correct Answers/Number of Questions* (9)	3 ADT fragments with 3 Y/N questions each
	Understandability in context Efficiency (UIC effc.)	*UIC Effectiveness/Time*	
	Transferability Effectiveness (TRF effv.)	$\sum$ *Correct Modifications/Number of Requests* (9)	3 ADT fragments with 3 requests each
	Transferability Efficiency (TRF effc.)	*TRF Effectiveness/Time*	
	Total Effectiveness	*Average of Understandability Dimensions Effectiveness*	Entire test
	Total Efficiency	*Average of Understandability Dimensions Efficiency*	
Acceptance	Perceived Ease of Use (PEOU) Perceived Usefulness (PU) Intention to Use (ITU)	Median of the questionnaire's statements points	Questionnaire (cf. Table 2)

Material. We measure the three variables through a questionnaire composed of a set of statements for each variable. We shuffle the statements and add their negated version to avoid systematic response bias (i.e., both the statements "ADTs are easy to learn" and "ADTs are not easy to learn" are present) [37]. Users need to evaluate each statement on a Likert scale from 1 (*strongly agree*) to 5 (*strongly disagree*). Table 2 shows the list of positive statements for PEOU, PU, and ITU. Each variable is computed as the median of its statements points (the points for negative statements are counted as 6 minus the points given as the answer).

Understandability. In line with the methodology outlined in [37], we tailor MEM by delineating specific objectives for evaluating ADTs: our study focuses on understandability. Understandability is evaluated in terms of effectiveness and efficiency based on the results of sample subjects in some problem-solving tasks (as suggested by the literature in, e.g., [69]).

For both effectiveness and efficiency, we further distinguish between fine-grained understandability, which considers three different dimensions of understandability separately, and coarse-grained understandability, which measures the average across the dimensions. The dimensions are:

UNC *Understandability not in context* measures the comprehensibility of ADTs *syntax*. It assesses users' ability, after ADT training, to identify correct ADT construction, to recognise nodes (for attack and defense), refinements (conjunctive and disjunctive), and countermeasures, and to understand sequential actions and their temporal order in ADTs.

Material. UNC is measured through a set of true/false questions on domain-agnostic ADT fragments (A, B, and C instead of names), to ensure that users' responses are not influenced by domain knowledge. Six items are presented, and for each of them, we show one or more ADT fragments and four statements regarding the syntax of the fragment(s). Participants have to check for each of the statements whether it is true or false, and they are asked to write down the starting (when starting this phase) and finishing time (when completing all the steps of this phase).

UIC *Understandability in context* measures the comprehensibility of the *semantics* of ADTs. It assesses users' ability, after training, to answer questions about both existing and instantiated ADTs and to recognise if an ADT accurately models a specific behaviour in a given scenario.

Material. UIC is evaluated through a set of yes/no questions on instantiated ADTs fragments. Three ADT fragments are presented, and, for each of them, a list of three yes/no questions regarding the semantics of the fragments. Participants are asked to answer the questions with "yes" or "no". The three ADT fragments used represent common and familiar types of attacks, namely an attack on a bank account, an attack to open a safe lock, and an attack to burgle a house. For each of the three items, participants are asked to write down starting and finishing times in the appropriate lines.

TRF *Transferability* measures the practical use of the notation, evaluating users' ability, after training, to create or modify ADTs. This includes recognising the appropriate elements to add to the tree for modelling specific behaviour and knowing where to place these elements.

Table 2
Perception-based statements (positive statements).

Variable		Statements
PEOU	1.	It was easy for me to understand what the ADTs represented.
	2.	ADTs are simple and easy to understand.
	3.	ADTs are easy to learn.
	4.	Overall, the ADTs were easy to use.
PU	1.	Overall, I think that ADTs provide an effective means for describing security threats and countermeasures.
	2.	I believe that ADTs have enough expressiveness to represent security threats and countermeasures.
	3.	Overall, I find ADTs to be useful.
	4.	I believe that ADTs are useful for representing security threats and countermeasures.
	5.	Using ADTs would improve my performance in describing security threats and countermeasures.
	6.	I believe that ADTs are organised, clear, concise, and unambiguous.
	7.	I believe the use of ADTs would reduce the time required to represent security threats and countermeasures.
ITU	1.	If I were to work for a company in the future, I would use ADTs to specify security threats and countermeasures.
	2.	I intend to use ADTs in the future if given the opportunity.
	3.	I would recommend the use of ADTs to security practitioners.
	4.	It would be easy for me to become skilled in using ADTs.

Material. TRF is measured through a number of instantiated ADTs fragments to extend with a set of requests. Three ADT fragments are presented (a simplified version of the fragments used to evaluate UIC) and, for each of them, a list of three requests, each with increasing levels of difficulty: (i) participants are asked to add a node to the tree and specify the type of node and its position; (ii) participants are asked to add all the nodes necessary to model a given situation; (iii) participants are asked to modify the tree according to given syntactic and/or semantic constraints. Participants are asked to modify the tree fragments according to the requests. For each of the three items, participants are asked to write down starting and finishing times in the appropriate lines.

For each of these dimensions, we compute effectiveness as the number of correct answers over the number of questions and efficiency as effectiveness over time [37]. For what concerns total understandability, we compute *understandability effectiveness* as the mean of the effectiveness of the three dimensions and *understandability efficiency* as the mean of the efficiency of the three dimensions.

4.2. Hypotheses

To answer the RQs, we test a number of NULL hypotheses (cf. Table 3). The hypotheses associated with RQ1 are oriented to assess whether both coarse- and fine-grained understandability effectiveness and efficiency are significantly above a certain sufficiency threshold—cf. Section 4.4 for the threshold values. It should be noted that, for the sake of synthesis, the effectiveness variables are identified with i in the hypotheses formulas reported in the table. The efficiency variables are identified with y. The hypotheses associated with RQ2 assess the variables associated with acceptances and check if they are significantly higher than the neutral value of the Likert scale. The hypotheses for RQ3 assess the relationships among the acceptance variables. Finally, the hypotheses for RQ4 evaluate whether there is a significant relationship between understandability variables and acceptance ones. Similarly to RQ1, effectiveness and efficiency variables are referred to with i and y, respectively.

4.3. Phases

The study is structured into six phases. The phases are illustrated in Fig. 2. The original experiment was conducted entirely online, utilising various online platforms for all phases. In contrast, the replication was conducted in person, with each phase adapted for face-to-face interaction. Details on the methods and platforms used in both experiments are provided in Section 5.

Table 3
Hypotheses for each research question. The variable i stands for total effectiveness, UNC effectiveness, UIC effectiveness, and TRF effectiveness. The variable y stands for total efficiency, UNC efficiency, UIC efficiency, and TRF efficiency.

RQ1	$\mathbf{H}_0(i)$	Users are not sufficiently effective in understanding ADTs
	$\mathbf{H}_0(y)$	Users are not sufficiently efficient in understanding ADTs
RQ2	$\mathbf{H}_0$(PEOU)	ADTs are perceived as difficult to use
	$\mathbf{H}_0$(PU)	ADTs are perceived as not useful
	$\mathbf{H}_0$(ITU)	There is no intention to use the ADT in the future
RQ3	$\mathbf{H}_0$(PEOU-PU)	There is no relationship between perceived ease of use and perceived usefulness
	$\mathbf{H}_0$(PU-ITU)	There is no relationship between perceived usefulness and intention to use
	$\mathbf{H}_0$(PEOU-ITU)	There is no relationship between perceived ease of use and intention to use
RQ4	$\mathbf{H}_0(i$-PEOU)	There is no relationship between i and perceived ease of use
	$\mathbf{H}_0(i$-PU)	There is no relationship between i and perceived usefulness
	$\mathbf{H}_0(y$-PEOU)	There is no relationship between y and perceived ease of use
	$\mathbf{H}_0(y$-PU)	There is no relationship between y and perceived usefulness

Phase 1—Recruitment. The participants are contacted through a recruitment e-mail with all the information needed to perform the study.

Phase 2—Binding. To ensure anonymity, the participants are provided with a unique alphanumeric identifier. They are instructed to keep the identifier for the entire test.

Phase 3—Training. Before starting the test, the participants are provided with a training session on ADT notation, which includes all the necessary information to successfully complete the test.

Phase 4—Pre-test questionnaire. We ask the participants to fill out a questionnaire collecting information about gender, age, education, employment, work area, level of knowledge of ADTs, and education on ADTs. The participants have to mark the questionnaire with the identifier received during the binding phase (Phase 2).

Phase 5—Test. We ask the participants to fill out the test in all its steps. The test is composed of four steps:

i **Retention.** Retention measures the comprehension of the training material and the ability to retain knowledge from it. We use this step to keep in the participants' memory the concepts presented in the training phase that they will need during the test. The outcome of this step is not utilised in the calculation of understandability. In this step, a list of figures (i.e., all figures in the legend of Fig. 1) is presented and, for each figure, a table with two definition options. Participants are asked to mark the right definition for each figure.

ii **Understandability not in context.** With this step, we want to identify how understandable the syntax of the notation is for the participants.

iii **Transferability.** Transferability measures to what extent the knowledge acquired through the training material is transferable.

iv **Understandability in context.** With this step, we want to identify to what extent users are able to answer questions about given ADTs (namely about the semantics of ADT notation).

Phase 6—Post-test questionnaire. We ask participants to fill out a questionnaire containing 8 statements concerning perceived ease of use, 14 statements on perceived usefulness, and 8 statements concerning intention to use (cf. Table 2). We use this phase to measure the perception-based variables (i.e., PEOU, PU, and ITU). Participants have to mark the questionnaire with the identifier received during the binding phase (Phase 2).

4.4. Data analysis

The experimental study protocol containing the definition of the study phases, its rationale, as well as the data analysis process has been submitted to the ethical committee of the Italian National Research Council (CNR), which authorised the administration of the test (authorisation number 0053588/2022). To take part in the study, participants are asked to sign an informed consent for the processing of personal data.

To answer RQ1, we first checked the dataset for normality using the Kolmogorov–Smirnov test, which resulted in p-values below the 0.05 significance level, indicating a failure of normality for all variables [68]. Due to this, we employed a non-parametric test, the Wilcoxon signed-rank test, to determine whether effectiveness and efficiency (both coarse- and fine-grained) are significantly above the selected target values. For what concerns effectiveness, we selected a target value of 0.6, which corresponds to 60% of answers being correctly answered; above such threshold, we consider understandability effectiveness (and its dimensions) as sufficient, based on the academic grading standards in Italy [70]. Regarding efficiency, users are not bound by a specific time frame for the test phase, but allocating 40 min for the overall questionnaire (10 min for UNC, 5 min for UIC, and 25 min for TRF) is deemed sufficient for completing all phases. This estimation is based on the expertise of the authors ter Beek and Lluch Lafuente, who are ADT experts [31,71]. This duration accounts for the time required for reading and analysing questions, processing ADT fragments, providing accurate answers, and adapting to the platform used. Therefore, we select a threshold of 0.015 for total efficiency, and 0.06, 0.12, and 0.024 for UNC, UIC, and TRF, respectively. These thresholds are computed as 60% of the maximum efficiency (i.e., 1) over the estimated time. To compute the effect size, we use the rank-biserial correlation, which is appropriate for non-parametric tests, both for the one-sample and paired samples cases [72].

To answer RQ2, we also checked for normality for all variables and, as with RQ1, the data failed the normality checks [68]. Consequently, we applied a Wilcoxon signed-rank test to assess whether PEOU, PU, and ITU scores are significantly above the value of the Likert scale representing neutral perception (i.e., 3). To compute the effect size, we use Cliff's delta, which is suitable for non-parametric tests with ordinal data [73].

To answer RQ3, we fit a regression linear model between PEOU and PU, and between both PEOU and PU and ITU.

To address RQ4 and investigate the potential relationship between the effectiveness and efficiency of understandability (both coarse- and fine-grained) and users' perceptions of ADTs' easiness and utility, a linear regression model was used to probe the association between perception-based variables (PEOU and PU) and both coarse- and fine-grained understandability effectiveness and efficiency.

Additionally, to compare the results on understandability effectiveness and efficiency (both coarse- and fine-grained) between the two experiments, we apply the Exact Wilcoxon rank sum test (also known as the Mann–Whitney U test) to determine whether there is a significant difference between them. We employ a non-parametric test due to the failure of normality checks for all variables using the Kolmogorov–Smirnov test. Moreover, we utilise the exact method, which is more suitable for managing ties for small datasets. This method provides exact p-values without relying on asymptotic approximations, ensuring robustness and accuracy in our statistical analysis. This test allows for a robust comparison, providing valuable insights into potential differences in the performance of the users across the experiments.

5. Individual experiments

In this section, we describe the two experiments, in terms of participants, phases, and material. In both experiments the participants were selected through convenience sampling, based on their availability. The phases follow the main structure defined in Section 4.3. Here, we describe the specific execution details, as well as differences between the two experiments.

5.1. Original experiment

5.1.1. Participants

In total, 25 participants took part in the study: computer science students (11), Ph.D. students (1), and professors (2); researchers in the field of software engineering (4), formal methods (3), and security (4). Participants belong to Kennesaw State University, CNR, University of Pisa, and the Technical University of Denmark. They were of both genders (56% men, 40% women, 4% prefers not to answer), aged between 21 and 56 years old. We asked them to self-evaluate their knowledge of ADTs before the test on a 5-point scale from 1 (*no knowledge*) to 5 (*advanced*) and whether they knew similar notations. The results are reported in Figs. 4(a) and 4(b), respectively. A total of 80% of the participants did not receive any education on ADTs before the test; the remaining participants attended a university course, a seminar, or were self-educated.

5.1.2. Phases and resources

Since the original study was conducted entirely online, each phase was designed to allow participants to complete all tasks while maintaining their anonymity.

During the recruitment phase, each participant received an email containing instructions to complete the test, the consent form (to be signed and returned), and links to the external resources and materials used for the study. Specifically, they were provided with a link to an online spreadsheet, where they could find their identifier and the associated link to the test document. Participants were instructed to retain their identifier throughout the test, preserve the link to the test document for subsequent phases, and use incognito mode to protect their identity.

Via a link received via email, participants could access an online training video (https://youtu.be/KLIH-yultgI)). They were asked to use this resource preferably once before beginning the test. If they watched the video multiple times, they were asked to indicate this in the post-test questionnaire.

Before starting the test, participants were asked to fill out an online pre-test questionnaire whose link had been sent by e-mail during the recruiting phase.

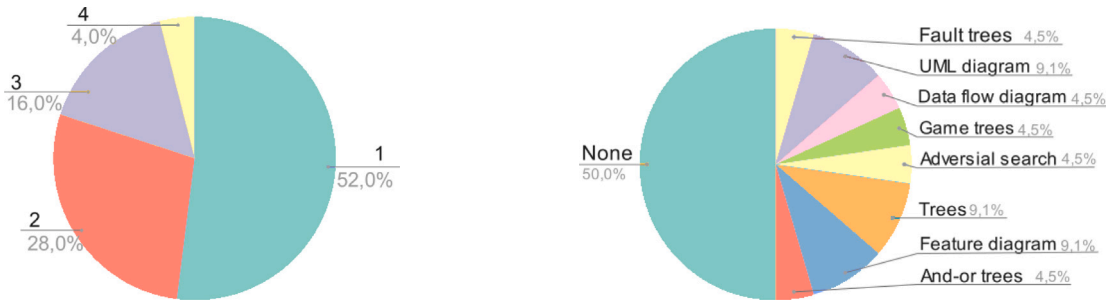
Subsequently, through the link received in the binding phase they could access to an editable online document (a different document for each participant). The spreadsheet accessed in the binding phase enables us to bind each document to the ID of the corresponding user. The online document contains the questions for all the test's sub-phases (cf. Section 4.3). The transferability step is conducted on an editable diagram embedded in the document (the instructions to modify the diagram are written inside the diagram itself).

After the test phase we asked participants to fill out the post-test questionnaire they could access through a link received in the recruiting e-mail.

5.2. Replication

5.2.1. Participants

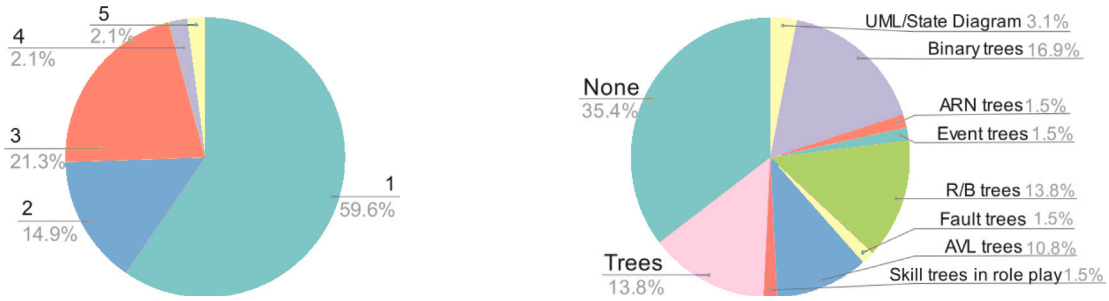
In total, 49 participants took part in the study. Two participants were excluded from the analysis: one due to missing data about the time of execution, and the other due to incompletion of the test. The participants were primarily computer engineering students attending



(a) Knowledge level of ADTs ranging from 1 (low) to 5 (high).

(b) Similar notations known.

Fig. 4. Participants' prior knowledge (original experiment).



(a) Knowledge level of ADTs ranging from 1 (low) to 5 (high).

(b) Similar notations known.

Fig. 5. Participants' prior knowledge (replication).

the course on industrial computing at the University of Florence, with the exception of one participant who was a master's student in modern Italian literature, and one participant who was a master's student in electronics engineering.

They were of both genders (87% men, 13% women), aged between 21 and 30 years old. The results of the self-evaluation about their knowledge of ADTs and about similar notations known are reported in Figs. 5(a) and 5(b), respectively. A total of 79% of the participants did not receive any education on ADTs before the test; the remaining participants attended a university course, a work course, or were self-educated.

5.2.2. Phases and materials

The replication was conducted in person. The motivation for the change of setting was due to the informal feedback received from some of the participants of the first experiment, who experienced difficulties with the online editable document used in the transferability phase. Since the replication of the study was conducted in person, all phases and materials were adapted accordingly to facilitate in-person management. The experiment was conducted in a university classroom at the University of Florence, supervised by two moderators (two of the authors of this article). Each participant received an experimental package containing the test to evaluate understandability, as well as the pre-test and post-test questionnaires, which were unified into a single document. All documents were marked with an identifier to maintain participant anonymity. Before the test began, participants attended a live training session that covered the same content as the video training used in the original experiment. Additionally, participants were provided with a link to the online video, allowing them to review the training if needed.

To ensure consistency with the original experiment, where participants had no interactions with the authors, participants in the replication were not allowed to ask questions during the test. This measure aimed to make the setup of the replication as similar as possible to the original experiment.

6. Results

Table 4 shows descriptive statistics for all the variables gathered with both the experiments (the original experiment and the replication), i.e., the perception-based variables (PEOU, PU, and ITU) and the performance-based variables: (1) understandability not in context effectiveness and (2) efficiency; (3) understandability in context effectiveness and (4) efficiency; (5) transferability effectiveness and (6) efficiency; and (7) understandability effectiveness and (8) efficiency.

Perception-based Variables. The medians for the perception-based variables are all above the neutral value of the Likert scale (i.e., 3), with a slight improvement in the replication. This suggests a general acceptance of the notation. Specifically, users perceive ADTs as easy to use, with median scores of 4.25 and 4.63 for the original study and the replication, respectively. They also find the notation useful, with median values of 4 and 4.14 for the original study and the replication, respectively. Furthermore, the results indicate that users intend to use the notation in the future, with ITU median scores of 3.88 and 4.25 for the original study and the replication, respectively.

The results indicate a generally good level of understandability for the ADT notation. The average total understandability effectiveness is 0.76 for the original experiment and 0.83 for the replication, both significantly above the sufficiency threshold of 0.6 (cf. Sections 6.1 and 6.2). Regarding total understandability efficiency, all averages exceed the sufficiency threshold, with better efficiency observed in the original experiment.

Table 4
Descriptive statistics. Columns denoted with *Or.* report the results for the original experiment; columns denoted with *Rep.* report the results for the replication.

Variables	Target	Median		Mean		Std. dev.		Min.		Max.	
		Or.	*Rep.*	*Or.*	*Rep.*	*Or.*	*Rep.*	*Or.*	*Rep.*	*Or.*	*Rep.*
PEOU	3	4.25	4.625	4.18	4.508	0.563	0.452	2.875	3.250	5	5
PU	3	4	4.143	3.92	4.131	0.37	0.536	2.929	2.929	5	4.571
ITU	3	3.875	4.250	3.88	4.176	0.403	0.588	3	2.750	5	5
UNC effectiveness	0.6	0.750	0.826	0.783	0.814	0.083	0.088	0.625	0.478	0.958	0.958
UNC efficiency	0.06	0.094	0.078	0.103	0.083	0.046	0.033	0.024	0.033	0.188	0.191
UIC effectiveness	0.6	0.889	1	0.907	0.922	0.175	0.095	0.111	0.667	1	1
UIC efficiency	0.12	0.250	0.222	0.264	0.228	0.135	0.084	0.009	0.089	0.500	0.444
TRF effectiveness	0.6	0.667	0.778	0.613	0.745	0.267	0.139	0	0.333	1	1
TRF efficiency	0.024	0.023	0.037	0.026	0.041	0.015	0.016	0	0.017	0.049	0.111
Understandability effectiveness	0.6	0.792	0.839	0.768	0.827	0.134	0.070	0.287	0.654	0.986	0.948
Understandability efficiency	0.015	0.118	0.113	0.131	0.117	0.059	0.31	0.011	0.060	0.241	0.194

Performance-based Variables. Regarding the different dimensions composing understandability effectiveness, the results show that understandability in context is the measure that provides the highest contribution (average effectiveness of 0.907 and 0.922 for the original experiment and replication, respectively), followed by understandability not in context (effectiveness of 0.783 and 0.814) and transferability (effectiveness of 0.613 and 0.745). Also in this case, we observe an enhancement in the performance during the replication. The results suggests that while participants understand the syntax and semantics of ADT fragments, they have more difficulty applying them in practice. For what concerns understandability efficiency, we observe a similar trend, thereby confirming that ADTs "in action" are perceived as more difficult.

Table 5 summarises the relation between variables expressed in the hypotheses addressing each RQ presented in Section 4.2. For each hypothesis, the column "*Rej.*" reports a "T" if the NULL hypothesis has been rejected and an "F" otherwise. Below we discuss in detail only the rejected NULL hypotheses because no conclusions can be drawn for the others.

6.1. Original experiment

RQ1. The test results show that all variables (both coarse- and fine-grained dimensions) are significantly higher than the target values for $\alpha = 0.05$ with large effect-size, with the exception of transferability effectiveness and efficiency (cf. Table 5). Figs. 6(a) and 6(c) show the boxplots illustrating the distribution of effectiveness and efficiency scores across total understandability, understandability in context, understandability not in context, and transferability. The plots show the central tendency and variability, with the median indicated by a horizontal line in each box. These results corroborate the descriptive statistics, suggesting a good level of understandability of the notation, particularly regarding the ADTs' syntax and semantics.

RQ2. As the boxplot in Fig. 6(e) shows, while ITU and PU have comparable values, PEOU receives the highest score. The test results show that all the variables attesting the acceptance are significantly higher than 3 for $\alpha = 0.05$, with p-values of 1.077e−05, 7.109e−06, and 9.282e−06, respectively, with large effect-size (cf. Table 5). This leads to the rejection of all three NULL hypotheses, confirming the overall high acceptance of the ADT notation.

RQ3. As shown in Fig. 6(b), the test results attest that there is a significant positive relationship between PU and ITU (p-value = 2.44e−06), suggesting that users are more likely to intend to use the notation in the future due to its perceived usefulness rather than its perceived ease of use.

RQ4. Regarding coarse-grained understandability, the test results reveal a significant positive relationship between effectiveness and perceived ease of use, and a weakly significant positive relationship between total effectiveness and perceived usefulness (cf. Fig. 6(d)). This suggests that users who perform better in the test tend to rate the notation higher in terms of easiness and usefulness.

For what concerns the understandability dimensions, our results show that understandability in context effectiveness and transferability effectiveness both have a significant positive relationship with PEOU, and both PEOU and PU, respectively. This suggests that users who observed instantiated trees and understand their meaning, tend to evaluate the notation as easier, whereas those who apply the method effectively by correctly extending the tree tend to find it both easier and more useful (cf. Fig. 6(f)). Thus, users who successfully use the notation in practice tend to appreciate it more.

6.2. Replication

RQ1. The results from the replication show that all variables measuring understandability are significantly higher than the target values, with large effect-size (cf. Table 5 and Figs. 7(a) and 7(c)). Unlike the original experiment, transferability is also significantly above the target values in this replication. This difference might be attributed to various factors, including the different support used (paper document instead of graphical editor online). Further analysis is required to clarify the reasons. Overall, these results confirm the previous findings regarding the good level of understandability of the notation.

RQ2. In the replication as well, the test results show that all the variables attesting the acceptance are significantly higher than 3 for $\alpha = 0.05$, with large effect-size (cf. Table 5 and Fig. 7(e)). Notably, ease of use shows an effect-size of 1, indicating a stronger and more significant result, thus suggesting that ease of use is the main characterising quality of ADTs.

RQ3. Regarding the relationship between the acceptance variables, all three NULL hypotheses have been rejected. This suggests that users who find the notation easier also tend to find it more useful and that they intend to use the notation in the future due to both its ease of use and its usefulness (cf. Figs. 7(d) and 7(b)).

RQ4. For what concerns the relation between understandability and the users perception about the notation's easiness and usefulness, the results of the replication contradict the original experiment's results. In this replication, the relationship between understandability not in context efficiency and perceived ease of use is significant (p-value = 0.002851), suggesting that users who understand the ADT syntax more efficiently tend to perceive the notation as easier (cf. Fig. 7(f)). Additionally, the relationship between transferability effectiveness and perceived usefulness is weakly significant, indicating that users who use ADTs more effectively tend to rate the notation higher in terms of usefulness. Overall, understandability does not seem to significantly influence users' perceptions.

6.3. Comparative analysis

To further enrich our analysis of the understandability of the ADT notation, we conducted a comparative analysis to assess the differences in effectiveness and efficiency between the original experiment and the

Table 5
Statistical test results addressing the research questions and their corresponding NULL hypotheses. For each hypothesis, the column *Rej.* reports a T if the NULL hypothesis has been rejected. Cells corresponding to rejected NULL hypotheses are highlighted in grey.

RQs	Hypothesis	Variable	Original experiment			Replication		
			p-value	Effect-size	*Rej.*	p-value	Effect-size	*Rej*
RQ1	H_0 (total effv.)	Total effectiveness	0.000139**	0.7291773	T	1.223e−09**	0.8705715	T
	H_0 (total effc.)	Total efficiency	7.381e−06**	0.8690932	T	1.233e−09**	0.8705715	T
	H_0 (UNC effv.)	UNC effectiveness	6.017e−06**	0.8744746	T	1.357e−09**	0.8659408	T
	H_0 (UNC effc.)	UNC efficiency	0.0001476**	0.7264866	T	7.145e−06**	0.633634	T
	H_0 (TRF effv.)	TRF effectiveness	0.1121	0.2448529	F	9.129e−08**	0.754804	T
	H_0 (TRF effc.)	TRF efficiency	0.2295	0.1506787	F	5.011e−09**	0.7123559	T
	H_0 (UIC effv.)	UIC effectiveness	7.603e−05**	0.7399401	T	5.424e−10**	0.8705715	T
	H_0 (UIC effc.)	UIC efficiency	4.191e−05**	0.7883725	T	3.932e−09**	0.8412437	T
RQ2	H_0 (PEOU)	PEOU	1.077e−05**	0.88	T	1.098e−09**	1	T
	H_0 (PU)	PU	7.109e−06**	0.92	T	1.984e−09**	0.9361702	T
	H_0 (ITU)	ITU	9.282e−06**	0.96	T	2.147e−09**	0.893617	T
RQs	Hypothesis	Relation between variables	Equation	p-value	*Rej.*	Equation	p-value	*Rej.*
RQ3	H_0 (PEOU-PU)	PEOU → PU	y = 3.6 + 0.073x	0.5962	F	y = 1.9 + 0.5x	0.003168**	T
	H_0 (PU-ITU)	PU → ITU	y = 2.9 + 0.24x	2.436e−06**	T	y = 1.3 + 0.63x	1.089e−07**	T
	H_0 (PEOU-ITU)	PEOU → ITU	y = 0.5 + 0.86x	0.108	F	y = 1.1 + 0.75x	0.0005451**	T
RQ4	H_0 (total ffv.-PEOU)	Total effectiveness → PEOU	y = 2.8 + 1.8x	0.03677**	T	y = 4 + 0.55x	0.5666	F
	H_0 (total effv.-PU)	Total effectiveness → PU	y = 3.2 + 0.97x	0.08483*	T	y = 3.4 + 0.84x	0.4647	F
	H_0 (total effc.-PEOU)	Total efficiency → PEOU	y = 3.8 + 2.7x	0.1752	F	y = 4.5 + 0.24x	0.912	F
	H_0 (total effc.-PU)	Total efficiency → PU	y = 4 − 0.25x	0.8492	F	y = 4.3 − 1.1x	0.6852	F
	H_0 (UNC effv.-PEOU)	UNC effectiveness → PEOU	y = 4.5 − 0.44x	0.7578	F	y = 3.8 + 0.83x	0.2799	F
	H_0 (UNC effv.-PU)	UNC effectiveness → PU	y = 4.5 − 0.74x	0.4241	F	y = 4.2 − 0.038x	0.9667	F
	H_0 (UNC effc.-PEOU)	UNC efficiency → PEOU	y = 3.9 + 2.9x	0.2606	F	y = 4 + 5.8x	0.002851**	T
	H_0 (UNC effc.-PU)	UNC efficiency → PU	y = 3.9 − 0.22x	0.8952	F	y = 3.8 + 3.6x	0.1357	F
	H_0 (UIC effv.-PEOU)	UIC effectiveness → PEOU	y = 2.8 + 1.5x	0.02051**	T	y = 4.8 − 0.36x	0.6096	F
	H_0 (UIC effv.-PU)	UIC effectiveness → PU	y = 3.5 + 0.43x	0.3332	F	y = 4.6 − 0.52x	0.5376	F
	H_0 (UIC effc.-PEOU)	UIC efficiency → PEOU	y = 3.9 + 1.1x	0.2168	F	y = 4.8 − 1.1x	0.1821	F
	H_0 (UIC effc.-PU)	UIC efficiency → PU	y = 4 − 0.16x	0.7812	F	y = 4.4 − 1.2x	0.2237	F
	H_0 (TRF effv.-PEOU)	TRF effectiveness → PEOU	y = 3.7 + 0.73x	0.08802*	T	y = 4.3 + 0.33x	0.5016	F
	H_0 (TRF effv.-PU)	TRF effectiveness → PU	y = 3.5 + 0.62x	0.02494**	T	y = 3.4 + 0.97x	0.08766*	T
	H_0 (TRF effc.-PEOU)	TRF efficiency → PEOU	y = 3.9 + 10x	0.2105	F	y = 4.2 + 6.5x	0.1147	F
	H_0 (TRF effc.-PU)	TRF efficiency → PU	y = 3.8 + 3.9x	0.4685	F	y = 4 + 4.2x	0.3968	F

* Indicates that results are weakly significant (*p-value* < 0.1).
** indicates that results are significant (*p-value* < 0.05).

Table 6
Statistics for comparative analysis.

Variables compared	p-value	Z
Understandability effectiveness	0.02722**	−2.2012
Understandability efficiency	0.3389	0.96419
UNC effectiveness	0.1523	−1.4377
UNC efficiency	0.09606	1.6681
UIC effectiveness	0.8909	−0.16962
UIC efficiency	0.2303	1.2072
TRF effectiveness	0.03155**	−2.1459
TRF efficiency	0.0006567**	−3.3381

** Indicates that the results are significant (*p-value* < 0.05).

replication for both coarse- and fine-grained understandability. This analysis aims to identify any significant variations and provide deeper insights into how the notation was perceived and utilised by participants in different settings. By examining these comparative results, we aim to draw more robust conclusions about the overall performance with the ADT notation.

Table 6 shows the results for all the tests. For what concerns coarse-grained understandability, there is a significant difference in understandability effectiveness between the original experiment and the replication (p-value = 0.02722), indicating that participants' effectiveness in understanding the ADT notation was significantly better in the replication (Z = −2.2012).

For what concerns fine-grained understandability, we can observe a weak significant difference in understandability not in context efficiency (p-value = 0.096), indicating that participants in the original experiment were more efficient in answering the questions on ADT syntax (Z = 1.6681). Moreover, we can observe a significant difference in transferability effectiveness and efficiency (p-values of 0.03155 and 0.0006567, respectively), indicating that participants in the replication were more effective and efficient in using the ADT notation in practice as compared to participants in the original experiment (Z values of −2.1459 and −3.3381, respectively).

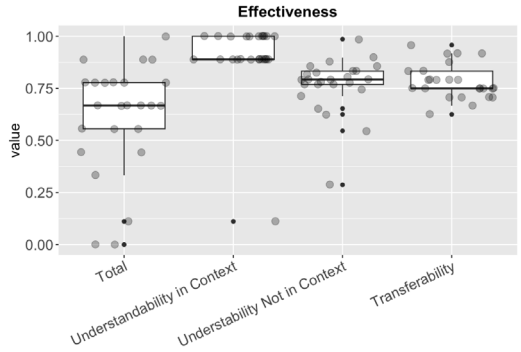
6.4. Discussion

The results from both the original experiment and the replication provide insightful observations on the understandability and acceptance of the ADT notation. This section will discuss common findings and highlight differences. Additionally, a discussion of the comparative analysis results is included.

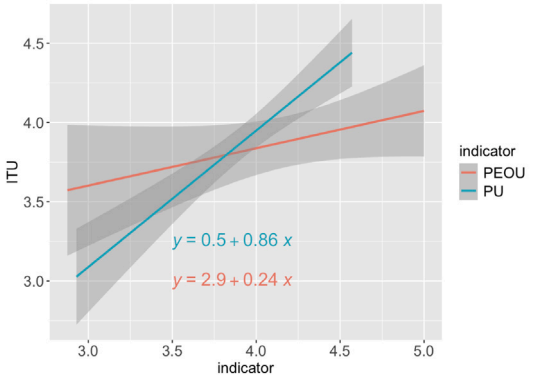
Common findings. Both studies demonstrate a generally high level of understandability for the ADT notation. All variables measuring understandability were significantly higher than the target values, suggesting that participants were able to grasp the ADT syntax and semantics effectively.

Concerning ADTs' acceptance, the results from both studies indicate that participants found the notation easy to use, useful, and intended to use it in the future. This is evidenced by the significant test results for all three variables (PEOU, PU, and ITU) in both studies, as well as the high median scores for each variable.

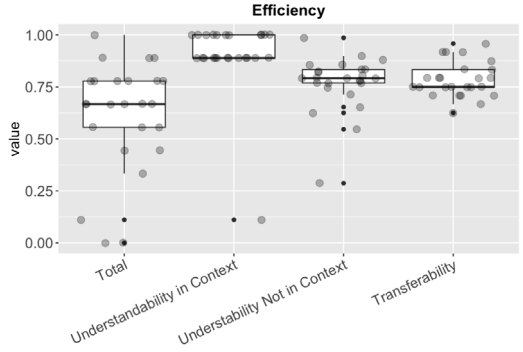
Additionally, the significant positive relationships identified in both studies between perceived usefulness and intention to use, suggest that users' willingness to adopt the notation in the future is largely driven by its perceived usefulness.



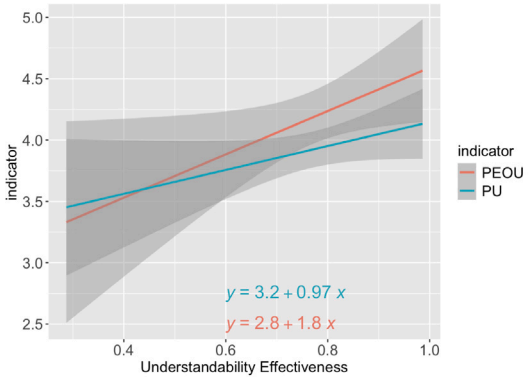
(a) Boxplots illustrating the distribution of effectiveness scores across four dimensions: total understandability, understandability in context, understandability not in context, and transferability.



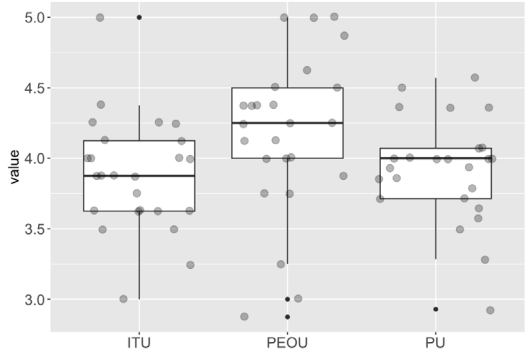
(b) Chart showing the relation between perceived ease of use (PEOU) and perceived usefulness (PU) on the x-axis, and intention to use (ITU) on the y-axis.



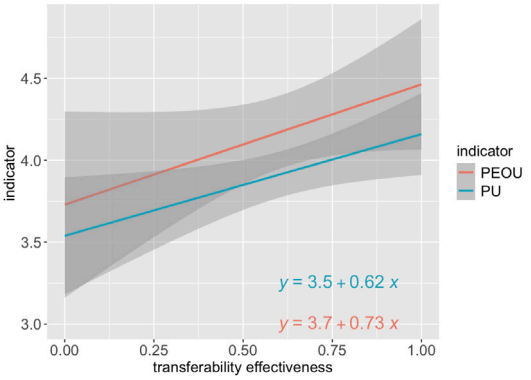
(c) Boxplots illustrating the distribution of efficiency scores across four dimensions: total understandability, understandability in context, understandability not in context, and transferability.



(d) Chart showing the relation between understability effectiveness on the x-axis and perceived ease of use (PEOU) and perceived usefulness (PU) on the y-axis.



(e) Boxplots illustrating the distribution of user acceptance scores across three dimensions: intention to use (ITU), perceived ease of use (PEOU), and perceived usefulness (PU).

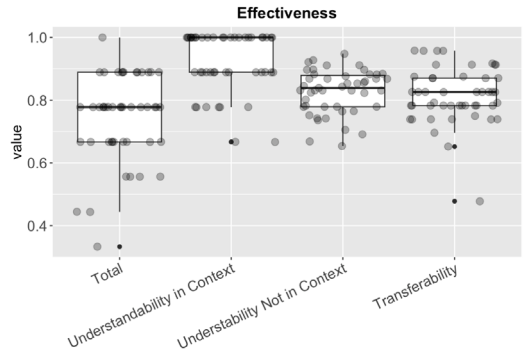


(f) Chart showing the relation between transferability effectiveness on the x-axis and perceived ease of use (PEOU) and perceived usefulness (PU) on the y-axis.

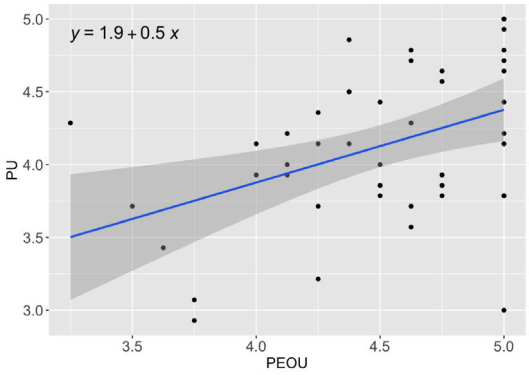
Fig. 6. Results charts of the original experiment.

Differences. Despite the overall consistency in findings, some differences were observed between the original experiment and the replication. One notable difference is in the effectiveness and efficiency of transferability, which were significantly higher than the target values in the replication. This could be attributed to the different supports used for conducting the experiments—online platforms for the original experiment and paper documents for the replication. Additionally, the in-person format of the replication might have provided participants with a more conducive environment for learning and applying the notation, thereby improving their performance in transferability tasks.

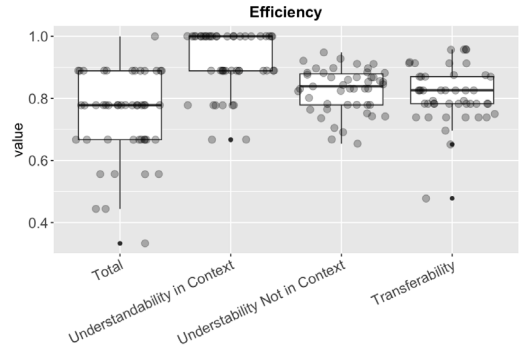
Moreover, the relationship between understandability and users' perceptions differed between the two studies. The only common result regards the relationship between transferability effectiveness and perceived usefulness, which was weakly significant in the replication. Due to the general inconsistency in the results for RQ4, nothing conclusive can be stated about the relationship between understandability and



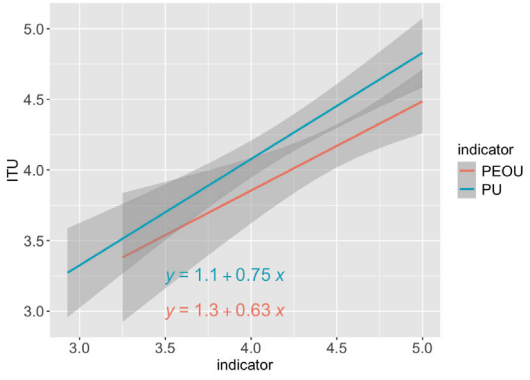
(a) Boxplots illustrating the distribution of effectiveness scores across four dimensions: total understandability, understandability in context, understandability not in context, and transferability.



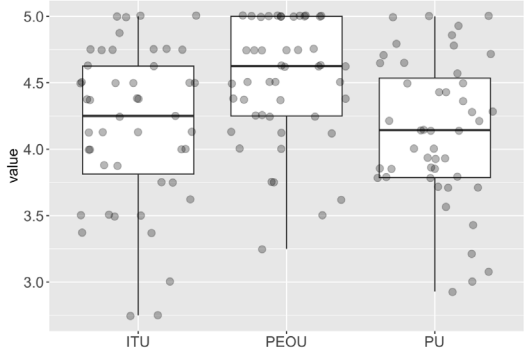
(b) Chart illustrating the relation between perceived ease of use (PEOU) on the x-axis and perceived usefulness (PU) on the y-axis.



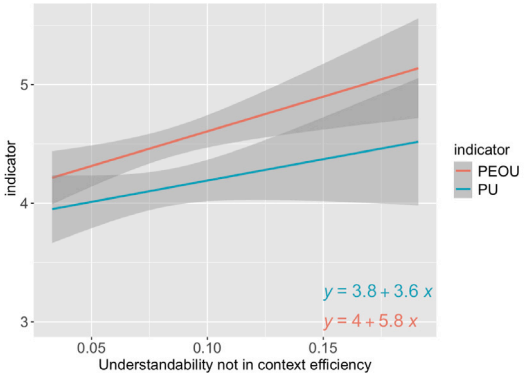
(c) Boxplots illustrating the distribution of efficiency scores across four dimensions: total understandability, understandability in context, understandability not in context, and transferability.



(d) Chart showing the relation between perceived ease of use (PEOU) and perceived usefulness (PU) on the x-axis, and intention to use (ITU) on the y-axis.



(e) Boxplots illustrating the distribution of user acceptance scores across three dimensions: intention to use (ITU), perceived ease of use (PEOU), and perceived usefulness (PU).



(f) Chart showing the relation between understandability not in context efficiency on the x-axis and perceived ease of use (PEOU) and perceived usefulness (PU) on the y-axis.

Fig. 7. Results charts of the replication.

acceptance. Different results are obtained in other studies, e.g. in [37], where a positive relationship between the two constructs appears to hold. In our case, we could hypothesise that users appreciate the notation independently from their performance. These uncertainties underscore the necessity for further investigation to comprehensively understand how users' perceptions of the ADT notation's ease of use and usefulness correlate with their actual performance in understanding and applying the notation.

More generally, the differences between the original experiment and the replication results can be attributed to several factors. Further investigation is needed to understand these differences comprehensively and to identify the underlying causes.

Comparative analysis. Regarding the comparative analysis, a significant difference in the coarse-grained effectiveness was noted, indicating an improvement in users' performance in the replication. While the replication group showed a higher level of understandability effectiveness, both groups demonstrated similar levels of efficiency in understanding the notation. The difference in effectiveness but not efficiency might be attributed to various factors such as differences in the study environment or the support used, and by the slight differences in terms of previous knowledge of similar notations. We can exclude the influence of the previous knowledge of the ADT notation, as this is basically equivalent – and very low – for both groups (cf. Figs. 4 and 5). The results on the fine-grained dimensions showed a significant difference only in transferability performance between the two groups (both effectiveness and efficiency). This suggests that the support used – embedded online diagrams vs. manual modification on paper documents – could have influenced participants' performance. It is possible that the paper documents were perceived as more suitable for applying the notation in practice compared to the online platform.

These findings may have practical implications for the development of ADT tools, emphasising the importance of providing clear and intuitive interfaces for users to interact with the notation, with the possibility that interaction similar to drawing may enhance user experience. The tool could potentially integrate manual sketching and modelling activities into the overall software engineering process, similar to what is done, (e.g., by FlexiSketch [74]). Furthermore, in absence of an enhanced ADT tool, users can perform the analysis with pen-and-paper, and then report it on one of the ADT tools available (e.g., ADTool, AttackTree), to facilitate sharing and collaborative modifications, when these features are available in the specific tool chosen.

Additionally, the observed differences highlight the need for further research to explore the impact of different study environments and supports on users' understanding and acceptance of the ADT notation, and graphical notations in general.

Another reason for the observed difference could be the knowledge of similar notations to ADTs. Participants who had already some confidence with other notations (replication study) could naturally be facilitated in the usage of the ADT notation in practice, compared to subjects with more limited knowledge of similar notations (original study). This has implications for teaching ADTs, as it suggests that this representation can be more successfully taught to students who have already used tree-like or similar modelling approaches.

Summary. In summary, both the original experiment and the replication confirm the high level of understandability and acceptance of the ADT notation among participants. The findings indicate that participants generally find the notation easy to use and useful, and that they are willing to use it in the future. Additionally, their performance in understanding and using the notation is overall good, demonstrating the effectiveness of the ADT notation. These findings confirm the assumptions of existing literature on the good understandability and ease of use of the ADT notation from the viewpoint of users [8]. The evidence reported in this paper can be exploited by tool developers and researchers to further promote the usage of the notation in industry.

In the comparative analysis between the two experiments, we have observed some differences especially for what concerns the practical application of the ADT technique, and we have argued that this can be associated with the platform used to support the activity—general purpose digital tool for drawing vs pen-and-paper. This provides suggestions for ADT tool developers, which should focus on easy to use paper-like interfaces to facilitate an effective exploitation of their tools.

7. Threats to validity

We present the threats to validity following the guidelines by Wohlin et al. [68], with a specific focus on construct validity, internal validity, and external validity.

Construct validity. Users' acceptance was assessed via existing models [22] and adapted to the ADT notation according to [37]. The usage of effectiveness and efficiency for understandability performance is widely used in the literature (cf., e.g., [22,37,39]). For what concerns the understandability dimensions, transferability is adapted from [37, 75]. Instead, understandability not in context and in context are measures adapted from [37] to address the evaluation of syntax and semantics. The problem-solving tasks proposed in the study mirror typical activities performed by ADT users, including understanding the meaning and syntax of trees, reading existing trees, and constructing or extending ADTs. All other possible tasks (e.g., evaluating the attack/defense cost with statistical model checking [31]) have not been considered, and different outcomes might be observed with different tasks.

Internal validity. To prevent systematic response bias in user acceptance questionnaires, we mixed positive and negative statements. The original experiment was conducted entirely online. While this setting created a more naturalistic environment, minimising biases introduced by participants' awareness of being observed and diminishing the Hawthorne Effect, the support used during the original test (e.g., the editable online document and diagrams) may have influenced users' performance. Indeed, we have observed a higher performance in the second experiment, which was conducted by means of pen-and-paper support. This difference could be due to the different support used. However, more targeted experiments should be conducted in which the support variable is isolated to better asses its influence. These experiments should also capture the difficulties possibly encountered by participants when using the online document and diagrams.

The presence of the experimenter during the replication could have intensified the potential for the Hawthorne Effect, as participants might have altered their behaviour knowing they are being observed. To mitigate this effect, the interaction between experimenters and participants has been limited to the training phase. To study these hypotheses, further investigation with users must be carried out to understand their opinions about the presence of a moderator.

The possibility of a more collaborative execution of the test in the in-person environment of the replication cannot be entirely ruled out. Despite participants appearing autonomous during the test, participants could have copied the solutions from each other. To mitigate this, we monitored the participants during the tests. In principle, this problem could have occurred also during the original experiment, since we could not control possible interactions between participants. However, we notice that none of the participants knew each other, and had different provenance, which minimises this threat. It is also essential to note that the online setup of the original experiment could have offered the possibility to better solve the presented tests. Specifically, in an online setting, participants may have access to additional support through other individuals or media, potentially aiding their understanding and performance. To mitigate this threat, the tasks were designed to require reasoning ability, rather than mere notions. Therefore, both environments have unique attributes that could impact the test results, and that we attempted to mitigate through different strategies.

Concerning instrumentation, the experimental material was carefully designed to faithfully instantiate the tasks. However, different experimenters could have designed different materials, possibly resulting in different outcomes. To mitigate this aspect, the material has been revised by two ADT experts and considered appropriate to evaluate the understandability of the notation. Concerning selection threats, the experiments involved volunteers, who could be more motivated with respect to a larger population. This threat could not be mitigated entirely.

External validity. The selected participants represent a diverse range of experience levels, which helps to enhance the generalisability of the findings. However, participants were opportunistically chosen from an academic background, with varying levels of seniority. This selection may not fully represent all potential ADT user groups, potentially

Table 7
Summary of the responses to each research question based on the results of the original experiment and the replication.

RQ1	*How well do novice users with no or minimal prior knowledge of the ADT notation understand ADTs?*	Users exhibit a good level of ADT understandability across both fine-grained and coarse-grained dimensions, with average values significantly above the predefined sufficiency thresholds, especially concerning its semantics.
RQ2	*What is the degree of acceptance of ADTs by novice users with no or minimal prior knowledge of the notation?*	Users demonstrate a great appreciation of the ADT notation, particularly for what concerns its perceived ease of use.
RQ3	*What is the relationship between ease of use/usefulness of the notation and the intention to use it in the future?*	Users are more prone to use the notation in the future thanks to its perceived usefulness.
RQ4	*What is the relationship between the ADT understandability and the users' perception of ADTs' ease of use and usefulness?*	Users more effective in using the ADT notation in practice tend to perceive it as more useful.

impacting the study results. For instance, practitioners with hands-on experience in security modelling and real-world decision-making may approach ADTs differently. Their practical insights could influence their understanding and acceptance of ADTs, leading to variations in perceived usefulness and effectiveness compared to the academic participants. Further research involving practitioners and users from other fields is necessary to confirm the applicability of these conclusions across broader user groups. While in the first experiment, the proportion of males and females was comparable, the second experiment included a higher proportion of males, which could have influenced the results. This threat could not be mitigated due to convenience sampling. However, given that the results of the two experiments are generally comparable, we can assume that gender did not affect the outcome.

Additionally, the academic educational background of the University of Florence, where theoretical aspects of software engineering and formal notations similar to ADTs are contained in courses of the curriculum, might have influenced the results of the replicated experiment. Exposure to related concepts may have made students more receptive or adept at understanding ADTs. However, it is worth noting that similar trends were observed in the original experiment, which included participants from diverse institutions and countries, such as Kennesaw State University (USA), CNR and University of Pisa (Italy), and the Technical University of Denmark (Denmark). This suggests that the findings may not be limited to a particular educational context.

It should also be noted that this study is a controlled experiment, which aims at maximising internal validity and does not evaluate ADT users in a realistic setting, where contextual factors play a relevant role. Several strategies proposed by Wieringa et al. [76] could be applied to address these limitations. For instance, *Lab-to-Field Generalisation* would involve testing ADTs in practical, real-world environments with practitioners to evaluate how well the findings translate beyond the academic context. This approach would help determine the robustness of the conclusions in diverse, real-world scenarios.

8. Conclusion and future work

In this paper, we presented the first empirical study to assess the quality of ADTs in terms of users' acceptance and understandability and its internal replication. Our evaluation measures how well the notation can be used in practice. In particular, our study focused on assessing users' perception variables that attest the notation appreciation in terms of ease of use, usefulness, and intention to use, and of performance variables that attest the degree of understandability of the notation in terms of effectiveness and efficiency. Understandability has also been studied according to three different fine-grained dimensions, and the relation between all these variables has been evaluated through multiple statistical tests.

Table 7 shows a summary of the responses to all RQs based on the results of both the original experiment and the replication. Our results suggest that the ADT notation is well understood and greatly appreciated by users; specifically, the main aspect characterising its quality is its ease of use. Overall, the notation has a good level of understandability with a total average effectiveness significantly above the sufficient threshold for both experiments. Among its dimensions, we note better performance in more practical tasks (i.e., those related to observing and extending instantiated trees). Concerning relationships among understandability and acceptance, we note a distinction between the two experiments making less clear the relationship between these two aspects.

In future research, to enhance result accuracy, we will broaden our subject pool, including users from diverse classes, such as those in the security field. We also intend to compare user performance and perceptions across ADTs and other security requirements modelling techniques, preferably textual methods. Additionally, our analysis will encompass various commercial and academic ADT tools. Finally, we plan to address user challenges in the test by conducting interviews to assess the impact of the platform on performance.

CRediT authorship contribution statement

Giovanna Broccia: Writing – original draft, Methodology, Data curation, Conceptualization. **Maurice H. ter Beek:** Writing – review & editing, Supervision, Project administration, Conceptualization. **Alberto Lluch Lafuente:** Writing – review & editing, Supervision. **Paola Spoletini:** Writing – review & editing. **Alessandro Fantechi:** Writing – review & editing. **Alessio Ferrari:** Writing – original draft, Supervision, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data available at [26].

References

[1] B. Fabian, S. Gürses, M. Heisel, T. Santen, H. Schmidt, A comparison of security requirements engineering methods, Requir. Eng. 15 (2010) 7–40, http://dx.doi.org/10.1007/s00766-009-0092-x.

[2] L. Liu, E.S.K. Yu, J. Mylopoulos, Secure-i*: Engineering secure software systems through social analysis, Int. J. Softw. Inform. 3 (1) (2009) 89–120.

[3] T. Lodderstedt, D.A. Basin, J. Doser, SecureUML: A UML-based modeling language for model-driven security, in: J. Jézéquel, H.H. mann, S. Cook (Eds.), Proceedings 5th International Conference on the Unified Modeling Language, UML, in: LNCS, Vol. 2460, Springer, 2002, pp. 426–441, http://dx.doi.org/10.1007/3-540-45800-X_33.

[4] E. Paja, F. Dalpiaz, P. Giorgini, Modelling and reasoning about security requirements in socio-technical systems, Data Knowl. Eng. 98 (2015) 123–143, http://dx.doi.org/10.1016/j.datak.2015.07.007.

[5] P.X. Mai, A. Goknil, L.K. Shar, F. Pastore, L.C. Briand, S. Shaame, Modeling security and privacy requirements: a use case-driven approach, Inf. Softw. Technol. 100 (2018) 165–182, http://dx.doi.org/10.1016/j.infsof.2018.04.007.

[6] B. Kordy, S. Mauw, S. Radomirović, P. Schweitzer, Foundations of attack-defense trees, in: P. Degano, S. Etalle, J.D. Guttman (Eds.), Revised Selected Papers 7th International Workshop on Formal Aspects of Security and Trust, FAST, in: LNCS, Vol. 6561, Springer, 2010, pp. 80–95, http://dx.doi.org/10.1007/978-3-642-19751-2_6.

[7] O. Gadyatskaya, R. Trujillo-Rasua, New directions in attack tree research: Catching up with industrial needs, in: P. Liu, S. Mauw, K. Stølen (Eds.), Revised Selected Papers 4th International Workshop on Graphical Models for Security, GraMSec, in: LNCS, Vol. 10744, Springer, 2017, pp. 115–126, http://dx.doi.org/10.1007/978-3-319-74860-3_9.

[8] W. Wideł, M. Audinot, B. Fila, S. Pinchinat, Beyond 2014: formal methods for attack tree-based security modeling, ACM Comput. Surv. 52 (4) (2019) 75:1–75:36, http://dx.doi.org/10.1145/3331524.

[9] J. Eisentraut, S. Holzer, K. Klioba, J. Křetínskỳ, L. Pin, A. Wagner, Assessing security of cryptocurrencies with attack-defense trees: Proof of concept and future directions, in: A. Cerone, P.C. Ölveczky (Eds.), Proceedings 18th International Colloquium on Theoretical Aspects of Computing, ICTAC, in: LNCS, Vol. 12819, Springer, 2021, pp. 214–234, http://dx.doi.org/10.1007/978-3-030-85315-0_13.

[10] H.S. Lallie, K. Debattista, J. Bal, An empirical evaluation of the effectiveness of attack graphs and fault trees in cyber-attack perception, IEEE Trans. Inf. Forensics Secur. 13 (5) (2018) 1110–1122, http://dx.doi.org/10.1109/TIFS.2017.2771238.

[11] H.S. Lallie, K. Debattista, J. Bal, A review of attack graph and attack tree visual syntax in cyber security, Comput. Sci. Rev. 35 (2020) http://dx.doi.org/10.1016/J.COSREV.2019.100219.

[12] Z. Sharafi, A. Marchetto, A. Susi, G. Antoniol, Y.-G. Guéhéneuc, An empirical study on the efficiency of graphical vs. textual representations in requirements comprehension, in: Proceedings 21st International Conference on Program Comprehension, ICPC, IEEE, 2013, pp. 33–42, http://dx.doi.org/10.1109/ICPC.2013.6613831.

[13] D. Stein, S. Hanenberg, R. Unland, A graphical notation to specify model queries for MDA transformations on UML models, in: U. Aßmann, M. Aksit, A. Rensink (Eds.), Revised Selected Papers European MDA Workshops: Foundations and Applications, MDAFA, in: LNCS, Vol. 3599, Springer, 2004, pp. 77–92, http://dx.doi.org/10.1007/11538097_6.

[14] K. Buyens, B. De Win, W. Joosen, Empirical and statistical analysis of risk analysis-driven techniques for threat management, in: Proceedings 2nd International Conference on Availability, Reliability and Security, ARES, IEEE, 2007, pp. 1034–1041, http://dx.doi.org/10.1109/ARES.2007.78.

[15] K. Labunets, F. Massacci, F. Paci, On the equivalence between graphical and tabular representations for security risk assessment, in: P. Grünbacher, A. Perini (Eds.), Proceedings 23rd International Working Conference on Requirements Engineering: Foundation for Software Quality, REFSQ, in: LNCS, Vol. 10153, Springer, 2017, pp. 191–208, http://dx.doi.org/10.1007/978-3-319-54045-0_15.

[16] K. Labunets, F. Massacci, F. Paci, L.M.S. Tran, An experimental comparison of two risk-based security methods, in: Proceedings 7th International Symposium on Empirical Software Engineering and Measurement, ESEM, IEEE, 2013, pp. 163–172, http://dx.doi.org/10.1109/ESEM.2013.29.

[17] F.D. Davis, Perceived usefulness, perceived ease of use, and user acceptance of information technology, MIS Q. 13 (3) (1989) 319–340, http://dx.doi.org/10.2307/249008.

[18] D.L. Moody, The method evaluation model: a theoretical model for validating information systems design methods, in: Proceedings 11th European Conference on Information Systems, ECIS, 2003, pp. 1327–1336, URL http://aisel.aisnet.org/ecis2003/79.

[19] V. Venkatesh, M.G. Morris, G.B. Davis, F.D. Davis, User acceptance of information technology: Toward a unified view, MIS Q. 27 (3) (2003) 425–478, http://dx.doi.org/10.2307/30036540.

[20] F.J. Shull, J.C. Carver, S. Vegas, N. Juristo, The role of replications in empirical software engineering, Empir. Softw. Eng. 13 (2008) 211–218, http://dx.doi.org/10.1007/s10664-008-9060-1.

[21] B. Kitchenham, The role of replications in empirical software engineering—a word of warning, Empir. Softw. Eng. 13 (2008) 219–221, http://dx.doi.org/10.1007/s10664-008-9061-0.

[22] D.L. Moody, Dealing with Complexity: A Practical Method for Representing Large Entity Relationship Models (Ph.D. thesis), University of Melbourne, 2001.

[23] G. Broccia, M.H. ter Beek, A. Lluch Lafuente, P. Spoletini, A. Ferrari, Assessing the understandability and acceptance of attack-defense trees for modelling security requirements, in: D. Méndez, A. Moreira (Eds.), Proceedings 30th International Working Conference on Requirements Engineering: Foundation for Software Quality, REFSQ, in: LNCS, Vol. 14588, Springer, 2024, pp. 39–56, http://dx.doi.org/10.1007/978-3-031-57327-9_3.

[24] A. Mavin, P. Wilkinson, S. Teufl, H. Femmer, J. Eckhardt, J. Mund, Does goal-oriented requirements engineering achieve its goal? in: Proceedings 25th International Conference on Requirements Engineering, RE, IEEE, 2017, pp. 174–183, http://dx.doi.org/10.1109/RE.2017.40.

[25] A. Ferrari, F. Mazzanti, D. Basile, M.H. ter Beek, Systematic evaluation and usability analysis of formal methods tools for railway signaling system design, IEEE Trans. Softw. Eng. 48 (11) (2022) 4675–4691, http://dx.doi.org/10.1109/TSE.2021.3124677.

[26] G. Broccia, M.H. ter Beek, A. Lluch Lafuente, P. Spoletini, A. Fantechi, A. Ferrari, Evaluating the understandability and user acceptance of attack-defense trees: a replicated experiment - replication package, 2024, http://dx.doi.org/10.5281/zenodo.11520632.

[27] W.E. Vesely, F.F. Goldberg, N.H. Roberts, D.F. Haasl, Fault Tree Handbook, Technical Report NUREG-0492, Nuclear Regulatory Commission, USA, 1981, URL https://www.osti.gov/biblio/5762464.

[28] B. Schneier, Attack trees, Dr. Dobb's J. (1999) URL https://www.schneier.com/academic/archives/1999/12/attack_trees.html.

[29] S. Mauw, M. Oostdijk, Foundations of attack trees, in: D. Won, S. Kim (Eds.), Revised Selected Papers 8th International Conference on Information Security and Cryptology, ICISC, in: LNCS, Vol. 3935, Springer, 2005, pp. 186–198, http://dx.doi.org/10.1007/11734727_17.

[30] B. Kordy, P. Kordy, S. Mauw, P. Schweitzer, ADTool: Security analysis with attack-defense trees, in: K. Joshi, M. Siegle, M. Stoelinga, P.R. D'Argenio (Eds.), Proceedings 10th International Conference on Quantitative Evaluation of Systems, QEST, in: LNCS, Vol. 8054, Springer, 2013, pp. 173–176, http://dx.doi.org/10.1007/978-3-642-40196-1_15.

[31] M.H. ter Beek, A. Legay, A. Lluch Lafuente, A. Vandin, Quantitative security risk modeling and analysis with RisQFLan, Comput. Secur. 109 (2021) 102381, http://dx.doi.org/10.1016/j.cose.2021.102381.

[32] B. Kordy, P. Kordy, Y. van den Boom, SPTool – equivalence checker for SAND attack trees, in: F. Cuppens, N. Cuppens, J.-L. Lanet, A. Legay (Eds.), Proceedings 11th International Conference on Risks and Security of Internet and Systems, CRiSIS, in: LNCS, Vol. 10158, Springer, 2016, pp. 105–113, http://dx.doi.org/10.1007/978-3-319-54876-0_8.

[33] R. Kumar, et al., Effective analysis of attack trees: A model-driven approach, in: A. Russo, A. Schürr (Eds.), Proceedings 21st International Conference on Fundamental Approaches To Software Engineering, FASE, in: LNCS, Vol. 10802, Springer, 2018, pp. 56–73, http://dx.doi.org/10.1007/978-3-319-89363-1_4.

[34] A.T. Limited, The SecuITree® BurgleHouse Tutorial (a.k.a., Who wants to be a Cat Burglar?), 2006, URL https://www.amenaza.com/downloads/docs/Tutorial.pdf.

[35] M. Audinot, S. Pinchinat, B. Kordy, Is my attack tree correct? in: S.N. Foley, D. Gollmann, E. Snekkenes (Eds.), Proceedings 22nd European Symposium on Research in Computer Security, ESORICS, in: LNCS, Vol. 10492, Springer, 2017, pp. 83–102, http://dx.doi.org/10.1007/978-3-319-66402-6_7.

[36] B. Kordy, W. Wideł, On quantitative analysis of attack-defense trees with repeated labels, in: L. Bauer, R. Küsters (Eds.), Proceedings 7th International Conference on Principles of Security and Trust, POST, in: LNCS, Vol. 10804, Springer, 2018, pp. 325–346, http://dx.doi.org/10.1007/978-3-319-89722-6_14.

[37] S. Abrahão, E. Insfrán, J.A. Carsí, M. Genero, Evaluating requirements modeling methods based on user perceptions: A family of experiments, Inform. Sci. 181 (16) (2011) 3356–3378, http://dx.doi.org/10.1016/j.ins.2011.04.005.

[38] K. Labunets, F. Paci, F. Massacci, M. Ragosta, B. Solhaug, A first empirical evaluation framework for security risk assessment methods in the ATM domain, in: Proceedings 4th SESAR Innovation Days, SID, EUROCONTROL, 2014, URL https://www.sesarju.eu/sites/default/files/documents/sid/2014/SID%202014-40.pdf.

[39] G. Broccia, A. Ferrari, M. ter Beek, W. Cazzola, L. Favalli, F. Bertolotti, Evaluating a language workbench: from working memory capacity to comprehension to acceptance, in: Proceedings 31st International Conference on Program Comprehension, ICPC, IEEE, 2023, pp. 54–58, http://dx.doi.org/10.1109/ICPC58990.2023.00017.

[40] D. Mellado, C. Blanco, L.E. Sanchez, E. Fernández-Medina, A systematic review of security requirements engineering, Comput. Stand. Interfaces 32 (4) (2010) 153–165, http://dx.doi.org/10.1016/j.csi.2010.01.006.

[41] I. Iankoulova, M. Daneva, Cloud computing security requirements: a systematic review, in: Proceedings 6th International Conference on Research Challenges in Information Science, RCIS, IEEE, 2012, pp. 1–7, http://dx.doi.org/10.1109/RCIS.2012.6240421.

[42] A. Souag, R. Mazo, C. Salinesi, I. Comyn-Wattiau, Reusable knowledge in security requirements engineering: a systematic mapping study, Requir. Eng. 21 (2016) 251–283, http://dx.doi.org/10.1007/s00766-015-0220-8.

[43] H. Villamizar, M. Kalinowski, M. Viana, D.M. Fernández, A systematic mapping study on security in agile requirements engineering, in: Proceedings 44th Euromicro Conference on Software EnGineering and Advanced Applications, SEAA, IEEE, 2018, pp. 454–461, http://dx.doi.org/10.1109/SEAA.2018.00080.

[44] M. Salehie, L. Pasquale, I. Omoronyia, R. Ali, B. Nuseibeh, Requirements-driven adaptive security: Protecting variable assets at runtime, in: Proceedings 20th International Requirements Engineering Conference, RE, IEEE, 2012, pp. 111–120, http://dx.doi.org/10.1109/RE.2012.6345794.

[45] G. Sindre, A.L. Opdahl, Eliciting security requirements with misuse cases, Requir. Eng. 10 (2005) 34–44, http://dx.doi.org/10.1007/s00766-004-0194-4.

[46] P. Giorgini, H. Mouratidis, N. Zannone, Modelling security and trust with secure Tropos, in: Integrating Security and Software Engineering: Advances and Future Visions, IGI Global, 2007, pp. 160–189, http://dx.doi.org/10.4018/978-1-59904-147-6.ch008.

[47] M.S. Lund, B. Solhaug, K. Stølen, A guided tour of the CORAS method, in: Model-Driven Risk Analysis: The CORAS Approach, Springer, 2011, pp. 23–43, http://dx.doi.org/10.1007/978-3-642-12323-8_3.

[48] M.S. Lund, B. Solhaug, K. Stølen, Model-Driven Risk Analysis: The CORAS Approach, Springer, 2010, http://dx.doi.org/10.1007/978-3-642-12323-8.

[49] F. Massacci, F. Paci, How to select a security requirements method? A comparative study with students and practitioners, in: A.J. sang, B. Carlsson (Eds.), Proceedings 17th Nordic Conference on Secure IT Systems, NordSec, in: LNCS, Vol. 7617, Springer, 2012, pp. 89–104, http://dx.doi.org/10.1007/978-3-642-34210-3_7.

[50] K. Labunets, F. Paci, F. Massacci, R. Ruprai, An experiment on comparing textual vs. visual industrial methods for security risk assessment, in: Proceedings 4th International Workshop on Empirical Requirements Engineering, EmpiRE, IEEE, 2014, pp. 28–35, http://dx.doi.org/10.1109/EmpiRE.2014.6890113.

[51] M. Jackson, Problem Frames: Analysing and Structuring Software Development Problems, Addison-Wesley, 2000.

[52] C. Haley, R. Laney, J. Moffett, B. Nuseibeh, Security requirements engineering: A framework for representation and analysis, IEEE Trans. Softw. Eng. 34 (1) (2008) 133–153, http://dx.doi.org/10.1109/TSE.2007.70754.

[53] European air Traffic Management (EATM), Security risk assessment methodology, 2008, pp. 42–43, URL https://www.eurocontrol.int/archive_download/all/node/11275.

[54] D. Mellado, E. Fernández-Medina, M. Piattini, Applying a security requirements engineering process, in: D. Gollmann, J. Meier, A. Sabelfeld (Eds.), Proceedings 11th European Symposium on Research in Computer Security, ESORICS, in: LNCS, Vol. 4189, Springer, 2006, pp. 192–206, http://dx.doi.org/10.1007/11863908_13.

[55] V. Volden-Freberg, G. Erdogan, An empirical study on the comprehensibility of graphical security risk models based on sequence diagrams, in: A. Zemmari, M. Mosbah, N. Cuppens-Boulahia, F. Cuppens (Eds.), Proceedings 13th International Conference on Risks and Security of Internet and Systems, CRiSIS, in: LNCS, Vol. 11391, Springer, 2018, pp. 1–17, http://dx.doi.org/10.1007/978-3-030-12143-3_1.

[56] M.S. Barik, C. Mazumdar, A graph data model for attack graph generation and analysis, in: G.M. Pérez, S.M. Thampi, R.K.L. Ko, L. Shu (Eds.), Proceedings 2nd International Conference on Recent Trends in Computer Networks and Distributed Systems Security, SNDS, in: CCIS, Vol. 420, Springer, 2014, pp. 239–250, http://dx.doi.org/10.1007/978-3-642-54525-2_22.

[57] I. Hogganvik, K. Stølen, A graphical approach to risk identification, motivated by empirical investigations, in: O. Nierstrasz, J. Whittle, D. Harel, G. Reggio (Eds.), Proceedings 9th International Conference on Model Driven Engineering Languages and Systems, MoDELS, in: LNCS, Vol. 4199, Springer, 2006, pp. 574–588, http://dx.doi.org/10.1007/11880240_40.

[58] A. Dikici, O. Türetken, O. Demirörs, Factors influencing the understandability of process models: A systematic literature review, Inf. Softw. Technol. 93 (2018) 112–129, http://dx.doi.org/10.1016/J.INFSOF.2017.09.001.

[59] T.R.G. Green, M. Petre, Usability analysis of visual programming environments: A 'cognitive dimensions' framework, J. Vis. Lang. Comput. 7 (2) (1996) 131–174, http://dx.doi.org/10.1006/JVLC.1996.0009.

[60] J. Sweller, Cognitive load during problem solving: Effects on learning, Cogn. Sci. 12 (2) (1988) 257–285, http://dx.doi.org/10.1207/S15516709COG1202_4.

[61] A. Alomary, J. Woollard, How is technology accepted by users? A review of technology acceptance models and theories, in: Proceedings 17th IRES International Conference, 2015, URL https://eprints.soton.ac.uk/382037/1/110-14486008271-4.pdf.

[62] V. Venkatesh, J.Y. Thong, X. Xu, Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology, MIS Q. 36 (1) (2012) 157–178, http://dx.doi.org/10.2307/41410412.

[63] M. Fishbein, I. Ajzen, Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research, Addison-Wesley, 1975.

[64] A. Bandura, Social Foundations of Thought & Action: A Social Cognitive Theory, Prentice Hall, 1986.

[65] I. Ajzen, The theory of planned behavior, Organ. Behav. Hum. Decis. Process. 50 (1991) 179–211, http://dx.doi.org/10.1016/0749-5978(91)90020-T.

[66] V. Venkatesh, H. Bala, Technology acceptance model 3 and a research agenda on interventions, Decis. Sci. 39 (2) (2008) 273–315, http://dx.doi.org/10.1111/j.1540-5915.2008.00192.x.

[67] F.J. Rondan-Cataluña, J. Arenas-Gaitán, P.E. Ramírez-Correa, A comparison of the different versions of popular technology acceptance models: A non-linear perspective, Kybernetes 44 (5) (2015) 788–805, http://dx.doi.org/10.1108/K-09-2014-0184.

[68] C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell, A. Wesslén, Experimentation in Software Engineering, second ed., Springer, 2024, http://dx.doi.org/10.1007/978-3-662-69306-3.

[69] D. Oliveira, R. Bruno, F. Madeiral, F. Castor, Evaluating code readability and legibility: An examination of human-centric studies, in: Proceedings 36th International Conference on Software Maintenance and Evolution, ICSME, IEEE, 2020, pp. 348–359, http://dx.doi.org/10.1109/ICSME46990.2020.00041.

[70] Wikipedia: The Free Encyclopedia, Academic grading in Italy, 2024, https://en.wikipedia.org/w/index.php?title=Academic_grading_in_Italy&oldid=1203203173. (Accessed 7 May 2024).

[71] M.H. ter Beek, A. Legay, A. Lluch Lafuente, A. Vandin, Variability meets security: Quantitative security modeling and analysis of highly customizable attack scenarios, in: Proceedings 14th International Working Conference on Variability Modelling of Software-Intensive Systems, VaMoS, ACM, 2020, pp. 11:1–11:9, http://dx.doi.org/10.1145/3377024.3377041.

[72] D.S. Kerby, The simple difference formula: an approach to teaching nonparametric correlation, Compr. Psychol. 3 (2014) 1:1–1:9, http://dx.doi.org/10.2466/11.IT.3.1.

[73] N. Cliff, Dominance statistics: Ordinal analyses to answer ordinal questions, Psychol. Bull. 114 (3) (1993) 494–509, http://dx.doi.org/10.1037/0033-2909.114.3.494.

[74] D. Wüest, N. Seyff, M. Glinz, FlexiSketch: a lightweight sketching and metamodeling approach for end-users, Softw. Syst. Model. 18 (2) (2019) 1513–1541, http://dx.doi.org/10.1007/S10270-017-0623-8.

[75] R.E. Mayer, Models for understanding, Rev. Educ. Res. 59 (1) (1989) 43–64, http://dx.doi.org/10.2307/1170446.

[76] R. Wieringa, M. Daneva, Six strategies for generalizing software engineering theories, Sci. Comput. Program. 101 (2015) 136–152, http://dx.doi.org/10.1016/J.SCICO.2014.11.013.