Transdisciplinary perspectives on validity: bridging the gap between design and implementation for technology-enhanced learning systems

Haastrecht, M.A.N. van

8

VAST: VALIDATING SOCIO-TECHNICAL SYSTEMS

The influx of technology in education has made it increasingly difficult to assess the validity of educational assessments. The field of information systems often ignores the social dimension during validation, whereas educational research neglects the technical dimensions of designed instruments. The inseparability of social and technical elements forms the bedrock of socio-technical systems. Therefore, the current lack of validation approaches that address both dimensions is a significant gap. We address this gap by introducing VAST: a validation framework for e-assessment solutions. Examples of such solutions are technology-enhanced learning systems and e-health applications. Using multi-grounded action research as our methodology, we investigate how we can synthesise existing knowledge from information systems and educational measurement to construct our validation framework. We develop an extensive user guideline complementing our framework and find through expert interviews that VAST facilitates a comprehensive, practical approach to validating e-assessment solutions.

The contents of this chapter are based on: van Haastrecht, M. J. S. Brinkhuis, Wools, et al. (2023). VAST: a practical validation framework for e-assessment solutions. Information Systems and e-Business Management.

8.1 INTRODUCTION

Educational assessments have to clear various hurdles before being used in practice. The test of validity is recognised as the most indispensable of these hurdles. Naturally, this has led to a flourishing discussion on validity theory and validation frameworks in the educational field. Regarding traditional forms of assessment, we have reached a point in the debate where most of the dust has settled. However, the influx of technology in education has altered the playing field. Technology introduces new possibilities for assessments, such as evaluating collaborative problem-solving skills (Stadler et al., 2020) and using learner behaviour analytics (Douglas et al., 2020). Yet, electronic assessments (e-assessments) also pose new challenges for validation. Tests can now be more interactive and complex (Mislevy, 2016), threatening our ability to judge validity due to decreasing transparency (Wools, Molenaar, et al., 2019). There is a need for e-assessment validation frameworks and that need is currently not catered to by the two fields from which we might expect a contribution: information systems (IS) and educational measurement.

The use of technology poses new questions regarding the validity of our tests but also necessitates a validity assessment of the technology itself. There is consensus in the IS field that a comprehensive evaluation is crucial when designing new artefacts (Hevner et al., 2004; Peffers et al., 2007). Action design research even considers the development and evaluation of an artefact to be inseparable (Sein et al., 2011). Nevertheless, the "discussion of evaluation activities and methods" remains limited (Pries-Heje et al., 2008) and current frameworks commonly offer "little or no guidance" to researchers performing evaluations (Venable et al., 2016). An inclination towards formulating general frameworks is a potential cause of the lack of guidance. Criteria that "can be applied to all research approaches" (Mingers and Standing, 2020) point to a focus on generality rather than specificity.

In educational measurement, where validation has been a central topic for nearly a century, the problem of open-ended validation approaches was a motivator for Kane (1992) to formulate argument-based validation. Subsequent work has recognised the usability of Kane's framework, but concurrently identifies areas where it lacks practicability (Cook et al., 2015). To solve this issue, Hopster-den Otter et al. (2019) traded generality for practicability. They introduced a validation framework for formative assessment contexts which offers clear guidelines to practitioners on how to use the framework.

Validation of complex systems stands to gain the most from practical frameworks such as that of Hopster-den Otter et al. (2019). Not only is the burden of proof high for complex systems, but researchers struggle to collect sufficient validity evidence for these systems due to their uncontrolled nature (Broniatowski and C. Tucker, 2017). A clear and transparent process for validation is crucial in such a situation.

Socio-technical systems (STS) are recognised for their tendency towards complexity. In STS, complexity arises from the number of components and the interactions between those components. Yet, we lack validation frameworks for STS. IS validation targets instrument validation as the core pursuit (Straub, 1989), essentially ignoring the social dimension. This is surprising when we consider that some researchers state that "information systems are socio-technical systems" (van Aken, 2013). Conversely, educational measurement validation focuses on the interpretation and use of an assessment by a learner, but avoids judging the validity of the technology. In this chapter, we take a first step in addressing this issue.

Given the progress in developing practical validation frameworks for formative assessment, it is worth investigating whether we can apply these insights to validate STS projects. Specifically, we focus on socio-technical solutions with assessment as a central aim: e-assessment solutions. Stödberg (2012) defines e-assessment to entail any assessment making use of information and communication technologies, where "the entire assessment process, from designing assignments to storing the results" is included. Examples of e-assessment solutions are technology-enhanced learning systems (M. J. S. Brinkhuis et al., 2018), e-health applications (Eskes et al., 2016), and cybersecurity risk assessment applications (van Haastrecht, Sarhan, Shojaifar, et al., 2021). With the need for a comprehensive, practical validation approach for e-assessment solutions in mind, we formulate the following research question:

- **RQ**: How can e-assessment solutions be validated comprehensively and practically?

In the remainder of this chapter, we will first provide the background to this work in Section 8.2. Section 8.3 covers the research methodology we applied in answering our research questions. In Section 8.4, we introduce VAST: the first comprehensive validation framework for e-assessment solutions. Section 8.5 presents the results of our grounding procedure, which centred around applying our validation framework in the EU cybersecurity risk assessment project GEIGER (GEIGER Consortium, 2020). The feedback we received inspired the development of an extensive user guideline to accompany the VAST framework (van Haastrecht, M. J. S. Brinkhuis, and Spruit, 2023). Where the VAST framework is the main theoretical contribution of this chapter, we envision the accompanying guideline to provide the most impact for practitioners striving to validate their solutions. We discuss the implications and limitations of our work in Section 8.6 and conclude in Section 8.7.

8.2 BACKGROUND

Thoughts on what constitutes validity have evolved over time and still differ across and within disciplines. In this section, we will cover those contributions

which help to understand the bigger picture of the validation literature, to create a common ground for the remainder of this chapter.

8.2.1 *Validation in educational measurement*

The field of educational measurement has close ties to psychological testing, which historically adopted a pluralist view on validity. Construct validity evolved from being an element in this pluralistic view to epitomising the overarching concept which unified all views on validity. A major proponent of this idea was Samuel Messick. Messick referred to the earlier pluralistic view as "fragmented and incomplete" and highlighted the need to integrate both "score meaning and social values in test interpretation and use" (Messick, 1995).

Although the validation framework Messick (1989) developed seemed to address many of the issues of earlier validation approaches, it was not very practical and "very open-ended" (Kane, 2013a). Kane (1992) introduced the argument-based approach to validation to address the open-ended nature of validation methods. Kane proposed a chain of inferences that form the interpretive argument, thereby giving guidance on "the kinds of inferences needed for the validation." Kane later extended this approach to an interpretation and use argument (IUA), aligning with the view of Messick that interpretation is not the only relevant dimension (Kane, 2004, 2013a).

Recent work has sought to provide guidelines on how to apply Kane's framework in particular contexts. Cook et al. (2015) provide a practical guide in the setting of medical education, noting that "Kane does not specify the order in which validity evidence should be collected and evaluated." Hopster-den Otter et al. (2019) extend the example inferences provided by Kane (2013b) for the context of formative assessment. The role of use is more prominent in formative assessment, which explains why Hopster-den Otter et al. (2019) chose to extend the IUA with additional use inferences. Although we have seen significant advances in the area of argument-based validation, Kane's framework has not yet been examined in assessment settings with a technological influx.

Modern times have seen the rise of technology-enhanced learning, with technology playing a part in our lives and education from an ever-younger age. With technology-enhanced learning becoming ubiquitous, one would expect an increased focus on validating e-assessments. Yet, although validating e-assessment requires specialised approaches (Wools, Molenaar, et al., 2019), no such approaches currently exist. This is a significant gap in the literature; a gap we aim to address in this work. To understand how we can best incorporate the technological viewpoint, we look towards the field that studies technological systems: information systems.

8.2.2 *Validation in information systems*

The seminal work of Straub (1989) on validity in IS outlines several validity types, as well as an order in which validation should address these types. Straub suggests to first conduct instrument validation, which consists of addressing content validity, construct validity, and reliability. Straub (1989) states that with content validity we answer the question: "Are instrument measures drawn from all possible measures of the properties under investigation?" This definition differs from the definition Cronbach and Meehl (1955) proposed in the educational measurement field, which states that test items should be an appropriate "sample of a universe in which the investigator is interested." Yet, the differences are somewhat superficial, as the underlying spirit is largely the same. Both definitions stress that content validity corresponds to how well we have sampled from the set of possible measurement items.

Straub's definition of construct validity also seems to depart from definitions as seen in Cronbach and Meehl (1955) and A. L. Brown and Campione (1996). Straub links construct validity to the question: "Do measures show stability across methodologies?" If stability is observed, we are dealing with valid constructs. Once more, however, the seeming disconnect with the more holistic definition of A. L. Brown and Campione (1996) is illusory. In later IS validation work based on Straub (1989), Mingers and Standing (2020) employ a definition which we feel strikes the right balance: "Do the measures converge on the construct and not on other distinct constructs?"

Reliability is the third element in Straub's instrument validation. Reliability answers the question of whether "measures show stability across the units of observation" (Straub, 1989). Although there is no direct analogue for this type of validity in the educational measurement field, inter-rater reliability is commonly incorporated in the inference chain of argument-based validation (Hopster-den Otter et al., 2019). Table 8.1 shows that Straub et al. (2004) mention Cohen's κ as a means of assessment for reliability. Cohen's κ is commonly used to measure inter-rater reliability.

Straub (1989) covers two further validity types: internal validity and statistical conclusion validity. Internal validity answers the question: "Are there untested rival hypotheses for the observed effects?" The underlying idea is that we should be confident in having identified the correct causal mechanisms at play in our setting. This is why we prefer to use the more direct definition of internal validity employed by Mingers and Standing (2020): "Are there alternative causal explanations for the observed data?" Statistical (conclusion) validity relates to the statistical robustness of validation results. If we can show that results are "unlikely to have occurred by chance" (Mingers and Standing, 2020), we add a further dimension to our overall validity claim.

Finally, Straub (1989) mentions the concept of external validity but states that "for the sake of brevity" it is not covered. In later work, Straub et al. (2004) link external validity to generalisability, but do not define the concept.

Table 8.1: Consolidated table of educational measurement and IS validity types considered in this chapter. Suggestions for means of assessment are provided for most validity types. Straub et al. (2004) and Mingers and Standing (2020) do not consider criterion validity and do not suggest means of assessment for internal and external validity.

TYPE	DEFINITION	MEANS OF ASSESSMENT
Construct validity	"Do the measures converge on the construct and not on other distinct constructs?" (Mingers and Standing, 2020)	Principal Component Analysis, Confirmatory Factor Analysis (Mingers and Standing, 2020; Straub et al., 2004)
Content validity	The extent to which measurement items are an appropriate sample from the universe of possible measurement items (Cronbach and Meehl, 1955; Straub, 1989).	Literature review, expert panel (Mingers and Standing, 2020; Straub et al., 2004)
Criterion validity	The extent to which test scores serving as an operationalisation of a construct correlate with an independent theoretical representation of the construct (i.e., the criterion) (Cronbach and Meehl, 1955).	Comparison to gold standard (Hopsterden Otter et al., 2019; Kane, 2013a)
External validity	"To what extent can the findings be generalised to other populations and settings?" (Mingers and Standing, 2020)	-
Internal validity	"Are there alternative causal explanations for the observed data?" (Mingers and Standing, 2020)	-
Reliability	"Do measures show stability across the units of observation?" (Straub, 1989)	Cronbach's α (Mingers and Standing, 2020; Straub et al., 2004), Cohen's κ (Straub et al., 2004)
Statistical validity	"Are the results sufficiently statistically robust that they are unlikely to have occurred by chance?" (Mingers and Standing, 2020)	R^2, F-test (Mingers and Standing, 2020), Structural Equation Modelling (Mingers and Standing, 2020; Straub et al., 2004)

Once more, we turn to the recent work of Mingers and Standing (2020) for our definition: "To what extent can the findings be generalised to other populations and settings?"

Criterion validity, a common concept in the educational measurement field, is largely ignored in the IS validation literature. We argue that in our context criterion validity is a vital element to consider alongside other validity types. This aligns with the prominent role Duolingo - the largest mobile language learning application - gives criterion validity in its validation approach. Duolingo's validity argument relies heavily on correlation with gold-standard language tests (Settles et al., 2020). Hence, we include criterion validity in our set of validity types presented in Table 8.1.

Since the work of Straub et al. (2004), the IS field has grown and changed considerably. The emergence of design science research saw the creation of new validation and evaluation frameworks. Work by Wieringa and Moralı (2012) and Venable et al. (2016) focused on suitable research methods for design science evaluation and validation. However, the initial focus Straub placed on instrument validity remained, meaning that the social element was still lacking in IS validation frameworks.

Frameworks linked to action research, such as that of Wieringa and Moralı (2012), more explicitly recognised the importance of the user. Yet, design science frameworks naturally target an evaluation of the designed artefact, rather than an assessment of validity. An example is the FEDS framework of Venable et al. (2016), which distinguishes the evaluation of purely technical

artefacts from the evaluation of artefacts involving a social component. This attention to social factors makes the framework more suited to STS, but an evaluation framework is not a validation framework. Where evaluation tends to focus on eliciting whether predefined performance indicators have been met, validation asks deeper questions on whether the designed artefact does what it was intended to do in its operational environment.

An additional problem is that current frameworks offer "little or no guidance" to researchers (Venable et al., 2016). The pluralistic view that is still dominant in IS validation today causes most frameworks to be complex and impractical. IS, like educational measurement, has not been able to solve the problem of open-ended validation. This is not a comforting thought when we consider that most IS validation frameworks do not recognise the social context of the instruments they are validating. We will require STS validation frameworks in the future and we need to avoid frameworks that are too general to be usable. Hence, we feel it is important to focus on the class of e-assessment solutions, where we can use insights from many decades of research in educational measurement validation to complement IS knowledge.

8.2.3 *Validation of e-assessment solutions*

In this section, we will cover three essential prerequisites for our validation framework: an existing validation framework to use as a basis, a modelling language to model e-assessment solutions, and an argumentation style for our argument-based validation approach. Regarding the first prerequisite, we use the Hopster-den Otter et al. (2019) formative assessment validation framework as the basis for our work. This framework extends the traditional IUA chain in argument-based validation with further inferences regarding use. The reasoning behind this extension is that a formative assessment validation framework must go beyond the inferences present in summative assessment frameworks. Formative assessment involves a translation of the outcome by the user to their situation, an evaluation of which actions they should take, and internalisation of the experience to learn.

Yet, the Hopster-den Otter et al. (2019) framework is not designed for STS. The terminology used (e.g., 'student learning') is specific to the classroom setting. To align the framework with STS, we draw on terminology from design research. Both educational and IS design research methods are employed when designing e-assessment solutions. Infusing the framework with terminology from these methods is our first step towards constructing an e-assessment validation framework. Figure 8.1 is the result of this process. The terms we introduce to the framework are inspired by the terminology used in the action design research work of Sein et al. (2011) and the educational design research work of McKenney and Reeves (2018).

Our second prerequisite is a modelling language to model the solution being validated. Any effort to validate an e-assessment solution must be

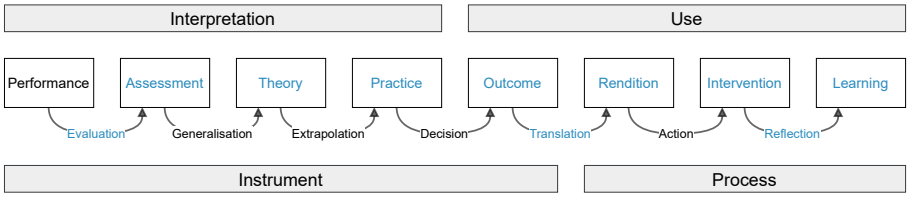

Figure 8.1: The inferences that make up the inference chain of the Hopster-den Otter et al. (2019) validation framework for formative assessment. Terminology that was adapted to suit our e-assessment setting is shown in blue.

predated by a description of that solution, consisting of the intended purpose and a representational model. We will assume that any researcher performing e-assessment validation has elicited functional and user requirements and is aware of the intended purpose of their system. This leaves the task of modelling the system.

Our STS model should, at minimum, include all relevant social and technical components and their interactions. If simplicity would not be a concern, flexible modelling languages such as Business Process Model and Notation (BPMN) and the Unified Modelling Language (UML) would be an ideal fit. However, BPMN and UML are notoriously complex modelling languages (Recker et al., 2009).

We should additionally acknowledge that we can treat the interpretation inferences of Figure 8.1 as being temporally independent, but that the same is not true for the use inferences. Use inferences depend on the thoughts and actions of users, which have a temporal structure. Hence, to address these inferences we must have a temporal model of our e-assessment solution. Finally, when evaluating use inferences it is preferable to initiate our argumentation from the user's perspective.

We have discerned that we require a modelling language that is not too complex, that allows for temporal dependencies, and that is user-oriented. We postulate that the answer lies in the use of user journey models. Any user journey representation that models all elements of an STS and their interactions satisfies the requirements we have put forth in this section. User journeys are temporal and user-oriented by nature. Therefore, a user journey modelling language that is not too complex can serve as the basis for our validation efforts. In this chapter, we employ the Customer Journey Modelling Language (CJML) (Halvorsrud et al., 2016; SINTEF Digital, 2022).

CJML models consist of temporally chained actions per actor. When an action constitutes an interaction with another actor in the system, CJML refers to this as a 'touchpoint.' Interactions have an initiator and a receiver. When multiple actors are involved, each actor has their own 'swimlane' in the CJML model. The corresponding diagram is termed a 'swimlane diagram.' The CJML swimlane diagram is the model we use in our validation framework.

Figure 8.2 shows an example swimlane diagram in the e-assessment setting. The user guideline (van Haastrecht, M. J. S. Brinkhuis, and Spruit, 2023) that accompanies this chapter contains several examples detailing how to construct a CJML diagram.

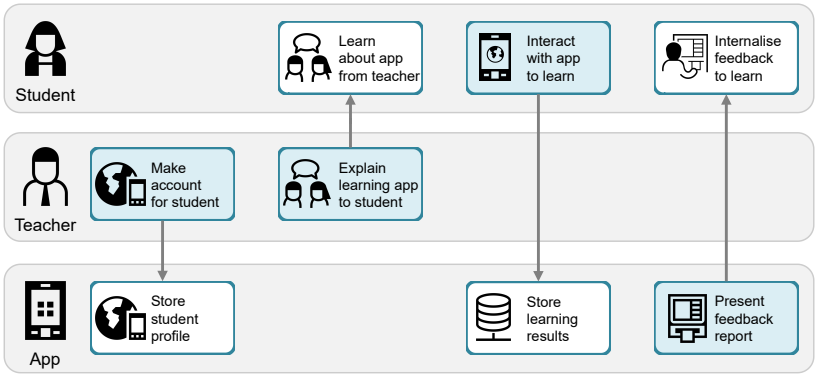

Figure 8.2: An example CJML swimlane diagram, where a student starts to use a mobile learning application. Each element of the system has a lane where actions are included in chronological order. When two elements of the system interact, the action of the actor initiating the interaction is coloured blue.

To address the final prerequisite for our validation framework, we will briefly cover the argumentation style we use within our argument-based validation approach. We choose to focus on Toulmin arguments since this style is commonly used in argument-based validation (Simon, 2008; Wools, Eggen, et al., 2010). Stephen Toulmin, a philosopher, introduced this structured style which divides argumentation into six components: claim, data, warrant, qualifier, rebuttal, and backing (Toulmin, 1958). Figure 8.3 depicts a Toulmin argument for the example of an online English language test.

8.3 METHODOLOGY

To synthesise theories from validation, modelling, and argumentation we require a flexible research methodology. We should build on existing theories and infuse our theory with insights from empirical work. Grounded theory is a research methodology suited to theory development. In its original definition, it was described as "the discovery of theory from data" (Glaser and A. L. Strauss, 1967). Grounded theory involves coding incidents found in the data into progressive abstractions to arrive at a theory, where 'incidents' are the basic units of analysis or ideas (Baskerville and Pries-Heje, 1999), and 'coding' involves the analysis and categorisation of incidents (Glaser and A. L. Strauss, 1967).

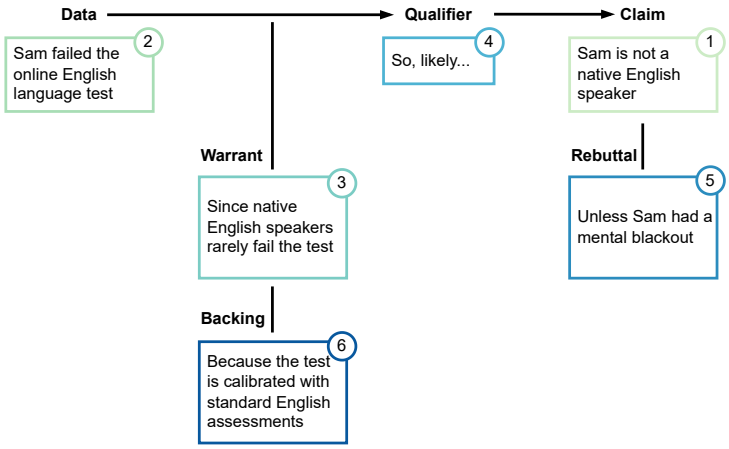

Figure 8.3: An example Toulmin argument for an online English language test. We want to make a claim (1) based on our data (2) and use a warrant (3) to support our claim. The qualifier (4) allows us to apply nuance to our claim. A rebuttal (5) can question the authority of our warrant, meaning we may require additional support to our warrant in the form of a backing (6).

Later extensions to grounded theory introduced three types of coding: open, axial, and selective (A. Strauss and Corbin, 1990). During open coding, the researcher aims to categorise essential incidents into concepts. Then, in axial coding, similar concepts are grouped into categories. Finally, selective coding works towards a core category, which from that point on is the main focus in the theorising process (Baskerville and Pries-Heje, 1999).

Grounded theory takes a purely inductive approach to theorising, meaning that in its strictest form grounded theory ignores established theories. The inductive approach has received heavy criticism, with some stating it constitutes a "loss of knowledge" (Goldkuhl and Cronholm, 2010). This led to the development of multi-grounded theory, where extant theories and knowledge receive a place in the theorising process. In multi-grounded theory, a researcher "constantly moves back and forward between data and preexisting knowledge or theories" (Thornberg, 2012).

Seeking to balance relevance-focused action research with rigour, Baskerville and Pries-Heje (1999) introduced the notion of grounded action research. The authors aimed for "a theory-rigorous and powerfully improved action research method," which remains practical and connected to organisational change (Baskerville and Pries-Heje, 1999). The multi-grounded variant of this approach soon emerged (Karlsson and Ågerfalk, 2007). Today, multi-grounded action research is positioned as the answer to how "knowledge development in action research [can] be clarified and improved" (Goldkuhl, Cronholm, and Lind, 2020). One way this manifests itself is in the three grounding approaches

present in multi-grounded action research: empirical grounding, theoretical grounding, and internal grounding. Emerging knowledge is grounded in empirical data through empirical grounding and in extant theories through theoretical grounding. Internal grounding helps to reflect on the emerging knowledge itself (Goldkuhl, Cronholm, and Lind, 2020). Figure 8.4 depicts the multi-grounded action research grounding procedure of our research. Extant theories contribute to the e-assessment validation framework through theoretical grounding and empirical data feeds into the emerging knowledge via empirical grounding. Lastly, expert evaluations provide internal grounding for our framework.

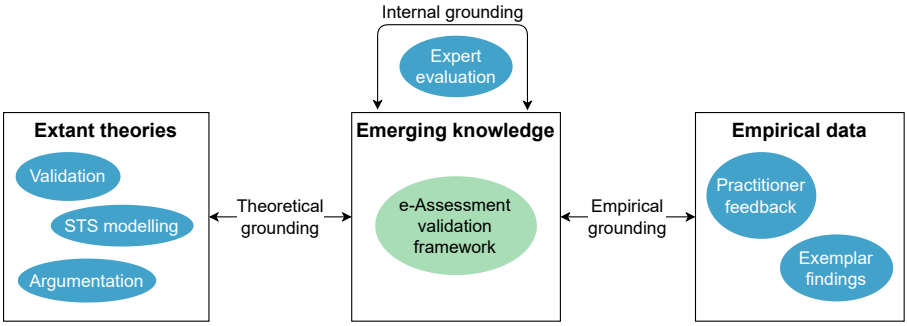

Figure 8.4: The grounding procedure of our multi-grounded action research methodology. Existing theories in validation, STS modelling, and argumentation provide theoretical grounding for our framework. We source empirical grounding from the practitioner feedback and exemplar findings that form our empirical data. Expert interviews help us to evaluate the internal cohesion of our emerging knowledge.

8.4 VAST

In this section, we propose VAST: an argument-based validation framework for e-assessment solutions. Traditional validation approaches consist of two main phases. First, a chain of claims specific to the project is constructed, which determines the inferences for which we need to provide arguments. Then, validity evidence is assembled to allow for a validity evaluation of our inference chain. However, in the complex setting of e-assessment, it is unclear where practitioners should source which evidence. VAST adds transparency to this process by inserting an additional step: modelling the system actions. The system model serves as a clarifying connector between the first and last steps in the validation process. Figure 8.5 presents the VAST framework. We will explain and motivate the three steps of VAST in the remainder of this section. For step-by-step instructions and practical examples of how to use the

VAST framework, we refer the reader to the VAST guideline (van Haastrecht, M. J. S. Brinkhuis, and Spruit, 2023).

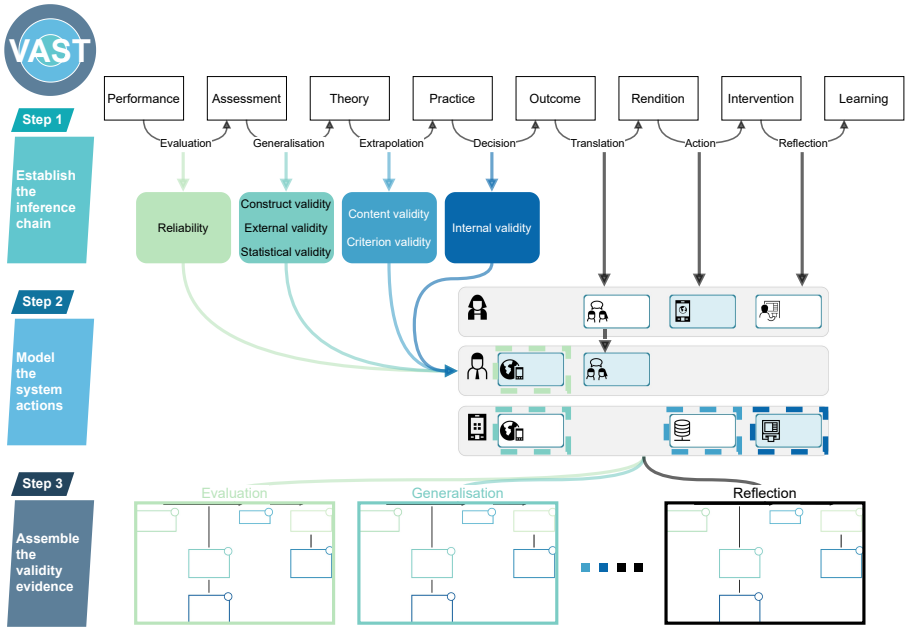

Figure 8.5: VAST: an argument-based validation framework for e-assessment solutions. VAST consists of three steps. Step 1 involves establishing the inference chain for the system being validated. By modelling the system actions in Step 2, we can match use inferences to user actions and instrument inferences to the remaining actions. Our model guides the assemblage of validity evidence in Step 3.

8.4.1 *Step 1: Establish the inference chain*

The first step within VAST consists of establishing the inference chain for the e-assessment solution at hand. We use our adapted version of the Hopster-den Otter et al. (2019) framework presented in Figure 8.1 as the starting point for this process. However, this is a general representation of an IUA chain, rather than a specific instance. Users of VAST will have to consider how the interpretation and use inferences materialise for their e-assessment solution. A vital prerequisite is that users have a clear idea of the objectives of their solution.

Part of this step will consist of making a first assessment of which inferences require more evidence than others. In certain systems, particular inferences will be redundant. As an example, consider the English language test we covered in Figure 8.3. If the test involves a diverse set of interactive written and

Table 8.2: Mapping of instrument inferences and validity types. We observe that the concepts related to instrument inferences relate to the concepts corresponding to particular validity types.

INFERENCE	CONCEPTS	VALIDITY TYPE	CONCEPTS
Evaluation	Consistency, inter-rater reliability	Reliability	Repeatability, stability
Generalisation	Theoretical constructs, different contexts, representative sample, control sampling error	Construct validity	Converge on construct
		External validity	Generalisation to other settings
		Statistical validity	Robust sample
Extrapolation	Accurate reflection of practice, theoretical tasks, compare to thorough assessments	Content validity	Appropriate sample, possible universe
		Criterion validity	Independent theoretical representation, comparison to gold standard
Decision	Underlying causal factors, outcome representation	Internal validity	Alternative causal explanations, observed data

oral exercises, the extrapolation inference taking us from theory to practice is largely obsolete. Although the option to prioritise inferences appears to introduce a layer of complexity to our framework, we want to stress that in principle all inferences should be considered. Only if a user of VAST is convinced that a particular inference is not relevant, should they disregard it.

Hopster-den Otter et al. (2019) connect their first four inferences to the instrument. In the following paragraphs, we will outline how we used our multi-grounded action research process to align the instrument inferences with IS validity theory. Table 8.2 shows the result of this work. In Section 8.4.2, we will investigate how we can synthesise the final three inferences with the e-assessment view.

In the evaluation inference, we assume that performance is consistently and reliably turned into assessment results. Inter-rater reliability is commonly mentioned as a possible source of evidence for this inference (Hopster-den Otter et al., 2019; Kane, 2013b). Mingers and Standing (2020) deem reliability to entail that results or responses are repeatable. This is similar to Straub (1989), who feels reliability should answer the question: "Do measures show stability across the units of observation?" We observe a clear connection between the concepts associated with the evaluation inference and with reliability. Hence, using the terminology of grounded theory, they are part of the same category.

In the generalisation inference, we assume the tasks of our assessment offer a sufficiently representative sample of the theoretical constructs we are aiming to represent (Hopster-den Otter et al., 2019). This ties the inference to our definition of construct validity outlined in Table 8.1. Additionally, it couples the inference to statistical validity. Statistical validity relates to whether our sampling approach is robust enough to rule out the possibility that results occurred by chance. This relates to the generalisation inference, which assumes that "tasks are sufficiently large to control sampling error" (Hopster-den Otter et al., 2019). We can observe from Table 8.2 that external validity relates to the generalisation inference. External validity addresses the following question: "To what extent can the findings be generalised to other

populations and settings?" (Mingers and Standing, 2020). This type of validity links to the generalisation inference, which extends the existing interpretation "to the expected performance over replications of the testing procedure (e.g., involving different test tasks, different testing contexts, different occasions, and raters)" (Kane, 2013a).

In the extrapolation inference, we assume that the theoretical tasks in the test domain accurately reflect practice (Hopster-den Otter et al., 2019). Content validity represents the extent to which test items are an appropriate "sample of a universe in which the investigator is interested" (Cronbach and Meehl, 1955). Content validity facilitates the extrapolation inference by motivating why our sample (theory) allows for an appropriate judgement regarding performance in the universe (practice) we are studying. A common way to support the extrapolation inference is to compare the results of our assessment to the results obtained by "assessments that cover the target domain more thoroughly" (Kane, 2013a). This corresponds to obtaining a gold standard result to compare to. This type of circular reasoning is both the link between criterion validity and the extrapolation inference and the "fundamental problem" (Kane, 2013a) of criterion validity.

The final inference we must account for is the decision inference, where a decision rule determines the outcome of our formative assessment. The choice of how to inform the user of the formative assessment outcome is vital, as it is the impetus for the formative process demarcated by the 'use' component of the IUA. This choice will be largely based on the causal factors that we assume to have generated the user's performance. With internal validity, we ask the question: "Are there alternative causal explanations for the observed data" (Mingers and Standing, 2020)? The internal validity of our e-assessment solution will determine whether we can formulate plausible backings for our decision inference. Hence, internal validity is the logical partner for the decision inference.

Our reasoning in the preceding paragraphs produced a coupling between the instrument inferences and the validity types of Table 8.1. The question remains how we can incorporate the inferences primarily related to use.

8.4.2 *Step 2: Model the system actions*

The second step in VAST consists of modelling the e-assessment system. We covered various STS modelling languages in Section 8.2.3, concluding that user journey modelling languages (specifically CJML) were best suited to our purpose. Figure 8.6 depicts the two stages involved in mapping the IUA inferences to our CJML model for the example covered in Figure 8.2. Recall that we are looking to inform the three use inferences: translation, action, and reflection. We posit that if any of the use inferences are of importance for an e-assessment solution, we can find a direct connection to at least one user

action corresponding to that inference. In our simple example of Figure 8.6, we see that each use inference connects to exactly one user action.

We connect the action of learning about the e-assessment application from a teacher to the translation inference. We reason that this introduction, whereby the teacher also learns from the student how they intend to use the application, will help in linking the eventual assessment to the student's circumstances. Given the inherent personal interactions that are present for the use inferences, we include the action of the teacher in this inference too. We denote this with a dotted, black arrow in Figure 8.6. If the user would have to perform additional actions themselves before the translation inference action, we would also connect these actions using dotted black arrows. Thus, we relate all actions to the translation inference that could directly or indirectly influence its interpretation in this context.

Similarly, we connect the action and reflection inferences to the CJML user actions. We connect the action inference to the interaction with the application and the reflection inference to the internalisation of feedback. Neither of these actions involves an interaction with another human actor. Rather, they constitute interactions with the application. Hence, we do not see any dotted arrows emanating from these actions.

Four actions remain unaccounted for. These are all the actions by actors that are not the student, except for those actions by human actors that involve direct interaction with the student. In a more general setting, we would refer to the student as the (main) social actor. Note that the actions that remain are not related to use, but rather to the instrument and preparatory work to enable later use. These are the actions that we can connect to the earlier inferences; the inferences regarding the interpretation and the instrument.

To couple the instrument inferences to the CJML diagram we can follow a more flexible approach. The instrument inferences do not need to abide by the temporal structure of the user journey model. Instead, we evaluate for each action which inference is most relevant. We circle the action using the colour of the most relevant inference. We see the result of this process in Figure 8.6. In our example, each inference corresponds to exactly one action. However, it is possible, and for larger e-assessment models often necessary, to map multiple actions to a single inference.

After completing the second step, the user of VAST will have gained further understanding of the system they are validating. Nevertheless, we have not yet assembled any validity evidence in the form of arguments. This is the focus of the third VAST step.

8.4.3 *Step 3: Assemble the validity evidence*

In the third and final step of VAST, the structured argumentation we discussed in Section 8.2.3 enters the stage. Figure 8.3 depicted the structure used in our arguments to motivate the inference from a datum to a claim. In the context of

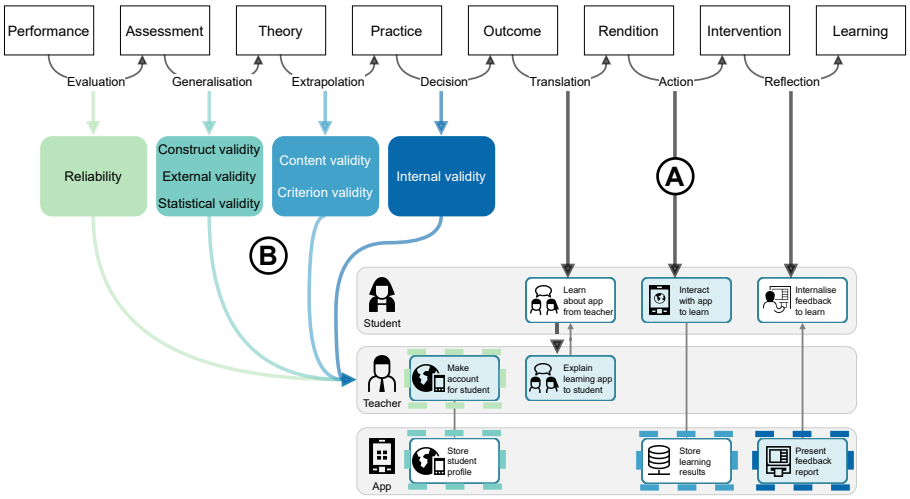

Figure 8.6: The two main stages involved in mapping the inference chain to the user journey model. First, we pair use inferences to user actions and interactions (A). The remaining actions are coupled to the inferences concerning the instrument (B).

our validation framework, we must provide argumentation for each inference, whereby the claim of the previous inference serves as the datum for the next inference.

Consider the evaluation inference, where we move from the performance datum to the assessment claim. We will at this stage have identified actions in our CJML diagram that relate to the evaluation inference and the corresponding validity type of reliability. Each action serves to inspire the relevant warrants, rebuttals, and backings that extend our argument. Although there is no absolute criterion to determine when an argument sufficiently motivates a claim, guidelines exist to assess the quality of argumentation. Erduran et al. (2004), for example, outline five levels of argumentation quality. From the lowest level 1 involving "a simple claim versus a counter-claim," we can improve to level 5 argumentation which "displays an extended argument with more than one rebuttal." Visually presenting the formulated arguments, as in the work of Wools, Eggen, et al. (2010), will then facilitate reviewers in assessing the quality of your argumentation.

Once we have provided sufficient evidence for the assessment claim, we proceed to the generalisation inference which connects the assessment datum to the theory claim. We continue along our inference chain until we have addressed all of our inferences. In this sense, the IUA and its argumentation serve "to specify what is being claimed" (Kane, 2013a). The final task is to assess the overall IUA with a validity argument, which "evaluates the plausibility of the proposed interpretations and uses" (Kane, 2013a). Kane

intends this to mean that the IUA is complete, coherent, and "supported by adequate evidence." VAST is structured to optimally address the validity argument.

Figure 8.5 depicts the three steps of establishing the inference chain, modelling the e-assessment solution, and assembling the validity evidence. Additionally, Figure 8.5 shows how the steps are connected to guide the user through the process. We believe the guidance provided within our framework allows us to counter the open-ended nature of validation and provide an actionable path towards validation. Although we have entrenched our framework in extant theories to provide theoretical grounding, we have not addressed the equally vital empirical and internal grounding within the multi-grounded action research grounding procedure. In Section 8.5, we turn our attention to practice to cover these further grounding procedures.

8.5 EVALUATING VAST

For our empirical and internal grounding, we applied VAST within the EU Horizon 2020 project GEIGER (GEIGER Consortium, 2020). GEIGER developed a cybersecurity risk assessment application for small businesses. The application helps raise employee awareness of cybersecurity threats and increase cybersecurity resilience. GEIGER assesses the cybersecurity risk faced by users and uses the outcome to offer personalised recommendations (van Haastrecht, Sarhan, Shojaifar, et al., 2021). By taking a formative approach to cybersecurity risk assessment, the GEIGER application forms an instance of the (formative) e-assessment solutions we are studying in this chapter.

To empirically ground VAST, we used an early variant of the framework to validate the GEIGER project. The details of this process are described in van Haastrecht, Spruit, et al. (2021). During six months of preparatory work, we gathered feedback on the first version of VAST from 13 different stakeholders across 14 sessions. We received comments that the framework did not offer enough practical guidance for validation. This feedback led us to include the second step of VAST, where the system is represented by a user journey model. The modelling step helped practitioners to connect abstract validation concepts to concrete user actions.

The updated version of our framework was further refined based on our first validation activities. These activities included an expert evaluation of the GEIGER content involving 14 stakeholders and user experience testing with our five use case partners (van Haastrecht, Spruit, et al., 2021). The findings from our practical application helped us to refine the step-wise approach of VAST, as it highlighted the necessity of forming a prioritisation among different validation activities. The refined variant of VAST was then further evaluated through interviews with validation experts.

To internally ground our framework, we interviewed three validation experts. We interviewed a senior researcher (SR) within the GEIGER project, an

Table 8.3: The current role, sector, and validation experience of our three interviewees. We use the ID to refer to the interviewees within the text.

ID	ROLE	SECTOR	VALIDATION EXPERIENCE
EA	External advisor	Private	15-20 years
PO	Project officer	Government	12 years
SR	Senior researcher	Academia	10 years

external advisor (EA) who is a member of the advisory board of GEIGER, and a European Commission representative with experience as a project officer (PO). All experts had at least ten years of validation experience at the time of the interview. Table 8.3 lists the details of the interviewees. The interviews consisted of a short introduction presentation explaining the GEIGER project and the VAST framework, followed by eight questions aimed at informing our internal grounding procedure. We list the interview questions in Table 8.4. Note that at the time of the interviews we had not yet developed the VAST guideline to accompany our framework.

With our main research question in mind, we asked the experts how VAST compared to traditional validation approaches regarding comprehensiveness and practicability. To make the concept of comprehensiveness more tractable for interviewees, we stated that this corresponds to coherence and completeness, using the terminology of Kane (2013a). EA and PO indicated that VAST would result in a much more coherent and complete validation process. SR stated that they could not compare VAST to earlier approaches in this way, since earlier approaches were always tailored to a specific project. Regarding practicability, EA and PO conveyed that VAST has the potential to at least be equally practical, given that users of the framework are well-prepared. SR suggested that more testing would be necessary to determine the practicability of VAST, although they too indicated that VAST has potential if it is supplemented with guidelines and practical examples on how to apply it.

Finally, EA and PO stated that they would likely recommend the use of VAST if they were to be involved in a future project of a similar nature. SR could imagine that they would recommend VAST given that adequate documentation and a practical, simple example of a VAST application exists. With internal grounding, we intend to investigate the "internal cohesion of the knowledge" being developed (Goldkuhl, 2004). Given the answers of our interviewees, VAST certainly exhibits internal cohesion. Nevertheless, there are areas for improvement, which we will cover in the following section.

8.6 DISCUSSION

Our grounding procedure demonstrated that although VAST helps to address the open-ended nature of validation, it cannot be considered a vali-

Table 8.4: Questions asked during the interviews with validation experts, in chronological order.

QUESTION	TYPE	OPTIONS
Please describe your previous validation experience (information systems, education, or other). How many years of experience do you have in your current role?	Open	-
What does validity constitute in your eyes? And validation?	Open	-
How do you view the original validation approach envisioned for GEIGER? Is it a similar approach to what you have encountered before?	Open	-
How do you view the VAST validation approach that was used for GEIGER? How appropriate do you think it is to build VAST on the argument-based formative assessment validation framework of Hopster-den Otter et al. (2019)?	Open	-
If you were to compare the VAST approach to the originally envisioned validation approach, how would you rate VAST in terms of coherence and completeness?	Likert	Much less, less, equal, more, much more
If you were to compare the VAST approach to the originally envisioned validation approach, how would you rate VAST in terms of practicability?	Likert	Much less, less, equal, more, much more
How likely are you to recommend the use of VAST for validation if you were to be involved in a future project of a similar nature?	Likert	Extremely unlikely, unlikely, neutral, likely, extremely likely
Is there anything else you would like to add? For example, something that you think can be improved in VAST.	Open	-

dation panacea. We have yet to see how VAST fares when applied to other e-assessment contexts, which themselves constitute only a fraction of all socio-technical systems. As we look to generalise, it is worth considering the observations of Addey et al. (2020). Though not outright disagreeing with the underlying push for clarity in Kane's argument-based validation, they observe that "in the quest for clarity and consensus, validity theory can become rarefied and idealised, and recognition of diversity diminished." Addey et al. (2020) note that Toulmin, who Kane builds on, shifted from an absolutist view on argumentation towards a more pluralistic one. Interestingly, this is in line with the view on validation we encounter in IS.

As we look to apply VAST in future work and generalise it to further socio-technical domains, we must always be wary of an overemphasis on clarity. We argued in our introduction that validation is inherently open-ended. When we take the pragmatic view of Kane too far, clarity becomes a requirement for successful validation, rather than a luxury. When this happens solutionism is just around the corner, especially in areas such as education where it already makes a regular appearance (McKenney and Reeves, 2021). Nevertheless, the reality of today's world is that complex systems exist and are continually being developed. As socio-technical systems increasingly become a part of our daily lives, we should not shy away from debating their validity. We believe frameworks such as VAST have a role to play in validation, as we strive towards clarity while recognising complexity.

Table 8.5 summarises the feedback we received in our expert interviews and the remarks of Addey et al. (2020) in three main suggestions for improvement of VAST. The first is to provide clarity where possible and appropriate. The experts we interviewed indicated that VAST would benefit from clear guidelines and supporting documentation, including practical examples of how to apply the framework. Focus groups could help us to improve the supporting

Table 8.5: The three axes of improvement identified for the VAST framework, resulting from a synthesis of the feedback from all interviewees. We briefly explain each concept and propose a possible research method we could use to investigate the implementation of each improvement.

IMPROVEMENT	EXPLANATION	RESEARCH METHOD
Clarify	Provide several practical examples on how to apply VAST, along with a clear step-by-step guideline and supporting documentation.	Focus groups
Modularise	Expand the scope of VAST outside of the e-assessment setting by providing custom inference chains for other STS classes.	Case studies
Visualise	Ensure that diverse perspectives on validity are recognised by designing a supporting tool where validity evidence can be assembled, debated, and visualised.	Educational design research

material in a collaborative, iterative fashion. To take a first step in addressing this axis of improvement and to signal our commitment to improving VAST, we created an extensive VAST guideline with practical examples to support future users (van Haastrecht, M. J. S. Brinkhuis, and Spruit, 2023). We plan to use focus groups to help us in iteratively refining the VAST guideline.

The second suggestion is to transform VAST to a more modular approach. By providing custom inference chains for other STS classes, we can expand the scope of VAST. A series of case studies could help to determine which STS classes, and corresponding inference chains, could be validated with a more flexible variant of VAST.

Finally, to ensure that VAST does not contribute to a diminishing recognition of diversity, we should develop a supporting tool which promotes a lively debate on validity. We agree with Addey et al. (2020) that argument-based validation needs "a democratic space in which legitimately diverse arguments and intentions can be recognised, considered, assembled and displayed." Following a design research methodology such as educational design research could be an appropriate approach to create such a tool.

8.7 CONCLUSION

Socio-technical systems are complex and difficult to validate, meaning we often have to rely on validity assessments that address only parts of the system. We investigated how e-assessment solutions, a particular class of socio-technical systems, can be validated comprehensively and practically. We compared and synthesised ideas regarding validation from the educational measurement and information systems fields. This resulted in an adaptation of the Hopster-den Otter et al. (2019) validation framework to suit the context of e-assessment.

We then used a multi-grounded action research approach to aid the development of VAST: an argument-based validation framework for e-assessment solutions. VAST is the first validation framework that explicitly combines validity theory from educational measurement and information systems. VAST

thereby addresses a significant gap that existed in the literature on socio-technical systems, namely the lack of validation approaches addressing both social and technical elements of the system being validated. We achieved this synthesis by identifying the commonalities between educational measurement inferences and information systems validity types.

Besides theoretical grounding, VAST resulted from empirical and internal grounding sourced from a practical implementation in the GEIGER project. We identified a need for clarity in the validation process, which VAST addresses by connecting inferences to concrete actions within the system. VAST additionally allows for transparent reporting of validation results by assembling validity evidence in the structure of Toulmin argumentation.

The validation experts we interviewed were assured of VAST's ability to facilitate a comprehensive and practical validation process. Still, the interviewees also provided suggestions for how to improve VAST. In future work, we hope to further VAST along the three axes of improvement identified by the experts: clarification, modularisation, and visualisation. We have already taken a first step in the area of clarification through the creation of an extensive VAST guideline containing practical examples (van Haastrecht, M. J. S. Brinkhuis, and Spruit, 2023). We expect this guideline, which incorporates concrete example use cases, to be of value to both researchers and practitioners. The foundation VAST provides spurs our confidence about the future of holistic socio-technical systems validation.