(In)validity by Design? Separating Measurement Validation from Substantive Research in HCI

Nele Borgert[1,*], Luisa Jansen[1] and Malte Elson[1]

[1]University of Bern, Institute of Psychology, Fabrikstrasse 8, 3012 Bern, Switzerland

Abstract

This position paper argues that the verity of empirical findings in HCI hinges on rigorous measurement practices of latent psychological constructs — a foundation too often undermined by ad-hoc and unvalidated scale development and modifications. We contend that current practices lead to distorted, only seemingly evidence-based inferences while squandering valuable resources. To mitigate these issues, we propose a 'validity by design' approach that separates measurement validation from substantive hypothesis evaluation. Central to our proposal is the establishment of a collaborative consortium tasked with defining core constructs, systematically validating standardized instruments, and disseminating them through open-access platforms. This coordinated effort — augmented by the active involvement of key thought leaders and targeted funding — will empower the HCI community to foster proactive measurement validation, thereby enhancing both scientific rigor and practical impact.

Keywords

measurement, construct validity, meta-science, human-computer interaction

1. Introduction

Measurement is fundamental to scientific progress. When researchers measure latent (non-observable) psychological variables — such as trust in technologies, attitudes, or self-efficacy beliefs — they must often rely on scales implemented in questionnaires that aim to translate abstract constructs into observable data. This challenge arises because latent constructs inherently lack "natural units of measurement" [1, p. 83], which can be resorted to when, for example, weighing kilograms. However, in many fields, including Human-Computer Interaction (HCI), it is common practice in quantitative studies to modify existing measures of latent constructs in an ad-hoc manner, adapting items, response formats, or scoring procedures within the same study in which a substantive hypothesis is tested [2]. This flexibility in measurement is rarely accompanied by systematic validation efforts, leading to an implicit assumption that 'minor' changes do not meaningfully affect the validity of interpretations based on the scale's scores. Yet, this assumption is in most cases both empirically unfounded and potentially harmful to the robustness of scientific conclusions [cf. 3].

This position paper advocates for the separation of construct validation efforts and the measures' implementation in studies with substantive research questions. In other words, measures must undergo cumulative (gradually built up) validation before being applied. It

Meta-HCI '25: First Workshop on Meta-Research in HCI, April 26, 2025, Yokohama, Japan

*Corresponding author.

nele.borgert@unibe.ch (N. Borgert); luisa.jansen@unibe.ch (L. Jansen); malte.elson@unibe.ch (M. Elson)

0000-0002-0255-5822 (N. Borgert); 0000-0001-8126-1306 (L. Jansen); 0000-0001-7806-9583 (M. Elson)

is the truth of scientific findings that depends on the validity of measured scale scores [4]. Without valid measures, research remains speculative and risks failing to generate meaningful recommendations for practitioners and policymakers.

2. Foundations of Construct Validity

When a score of a scale is interpreted as capturing a latent psychological attribute, construct validation becomes a necessary step to estimate the verity of this match [5]. A valid scale measures what it claims to measure [6]. Accordingly, the goal of construct validation is to provide evidence that the scale accurately maps onto the construct of interest; for example, higher scores on a scale can be meaningfully interpreted as higher levels of that specific attribute [7]. One of the main approaches in the psychological test literature is to engage with the so-called 'nomological network' of related constructs — postulated by theoretical works — and test whether empirical relationships align with the constructs' hypothesized connections. For instance, in [8], it was examined whether, in line with Bandura's Social Cognitive Theory [9, 10], the newly developed cybersecurity self-efficacy scale exhibited only marginal correlations with scales measuring self-esteem, outcome expectations, and optimism — constructs that the theory conceptualizes as distinct attributes.

Invalid measures have two major implications. First, a study's conclusions drawn from surveyed scores are unsound if the scale does not measure what it set out to measure [7]. This is because the reported statistical relationships lose their meaning: to illustrate, correlations between an invalid self-efficacy scale and behavioral outcomes do not yield useful insights about users' self-efficacy if, e.g., the scale in fact is more related to the optimistic attitudes of participants in general, rather than about their self-efficacy specifically. Second, using an invalid scale wastes valuable resources in the affected study (e.g., time, money, and participant contributions), and the problem extends beyond isolated cases. If multiple studies employ slightly altered versions of a scale without proper validation, meta-analyses become unreliable [11], replication efforts yield negligible benefits [7], and broader theoretical frameworks built on these measures are ultimately compromised. This amounts to a substantial and ongoing waste of the scientific community's collective efforts.

Hence, if researchers wish to introduce a new scale into the literature, the first step which is required is a systematic and in-depth approach to measurement validation. This is vital for interpreting a study's results and justifying the resource investment. Even if researchers merely adapt an existing measure of, for example, trust in technologies, but fail to confirm that it still reflects the construct as originally intended, they might be measuring something entirely different — such as usefulness or even intention to explore. This flexibility introduces uncertainty, and until validation data is provided "any conclusions are questionable" [4, p. 374]. It is the task of science to develop meaningful measures and defend their validity [12].

3. Undermined Construct Validity in HCI

To establish construct validity, a cumulative process of theoretical and empirical works is required, involving multiple types of validity evidence, including convergent and discriminant

validity [13]. Convergent validity refers to the extent to which a measure relates strongly with other measures of the same construct, while discriminant validity ensures that the measure is distinct from different and still reasonably related constructs [14]. When researchers modify a scale without renewed validation studies, they disrupt this process, introducing uncertainty about whether the new or adapted measure still accurately reflects the original construct. Every modification — no matter how minor — carries the risk of altering the postulated reflection of the construct. For example, if researchers remove or reword items of a scale to better align with their study's needs, they may unintentionally change what is being measured. Consequently, each altered version of a scale risks drifting away from its original meaning or, ideally, could be shown through validation to improve the overlap between scores and the construct of interest. Without separate validation, researchers have no way of determining whether their interpretations are based on true relationships [4]. More generally, if several studies concerning a similar substantive research question (e.g., whether trust in technologies influences technology adoption) employ slightly different, unvalidated versions of a scale, it becomes impossible to discern whether observed differences in findings reflect genuine theoretical variation or inconsistencies in measurement.

As researchers do not draw conclusions from participants' raw responses to items, but rather from claims about the psychological construct — claims that are of primary interest to researchers and are then prominently described in the scientific literature [12] — construct validity comes into focus. Relying solely on the name of a scale does not guarantee its validity, as a label does not necessarily reflect the scale's actual content or its alignment with the intended construct (cf. jingle-jangle fallacies) — let alone provide evidence for its validity [13]. Contrarily, theoretical arguments that clearly define the construct and empirical evidence that supports its position within the proposed nomological network, i.e., the theoretical relationships that determine how constructs operate and interact with others [5], are required. Without such a-priori evidence, any success or failure to accept a postulated substantive hypothesis leaves multiple points of uncertainty that could have caused the (un)expected results [5].

To make this explicit, if a field bypasses the prerequisite of validation when investigating latent constructs, scales may overfit to a single sample — capturing not only the intended construct but also unique characteristics of that group — thus potentially failing to generalize to other samples. The overfit can additionally be shaped by other questionable measurement practices [cf. 15] that may, either deliberately or accidentally, skew item wording, selection, or scoring toward a favored or 'beneficial' study outcome. Reported psychometric strengths of those measures may thus partly reflect noise and interpretation of results may be based on inflated effect sizes [cf. 16]. This would contribute to misleading recommendations for practitioners and policy-makers. Researchers need to be aware that unvalidated score interpretations are shaped by — and thus best suited for — the data from which they originate, resulting in a likely waste of measures that may be ineffective for further investigation of the construct. Therefore, robust measures necessitate a validation in independent samples and diverse contexts, while retaining standardized use. This problematic lack of scale reuse has been well documented across psychological research [17].

HCI faces these discussed risks, as measurement practices in the field have become fragmented and largely ad-hoc, often bypassing validation steps and thereby restricting the ability to investigate substantive and meta-analytic research questions. This has already been observed

in HCI for psychological constructs such as self-efficacy, [2], aggression [3], and trust [18]. We argue that current measurement practices in HCI implicitly build invalidity into research designs by routinely allowing unvalidated measures and ad-hoc modifications of measures without systematic revalidation — an 'invalidity by design' approach embedded within the research process itself. Each time researchers develop new or alter existing items, response formats, or scoring procedures without checking if the measure captures the intended construct, they embed uncertainty about the validity of interpretations directly into their studies from inception. Such implicit acceptance of measurement invalidity from the very start jeopardizes all potential meaning from subsequent research stages. It seems as though validity is assumed retrospectively, after substantive research questions have already been evaluated, creating a reactive cycle that perpetuates uncertainty and undermines the scientific rigor of the field.

The invalidity by design approach in HCI research may stem from a combination of tensions: (a) the field's emphasis on novel, context-specific technological solutions may compel researchers to rapidly adapt or develop scales to meet immediate project demands — such as different interaction modalities, interfaces, or user groups; and (b) scarce time and funding resources may force a prioritization of quick, high-volume research outputs, thereby encouraging 'pragmatic' shortcuts and the de-emphasis of issues that are not considered central to typical HCI publication outlets. On the other hand, achieving cross-situational standardization and systematic validation of measures requires a deliberate, long-term, and methodologically focused process that involves extensive empirical testing and theoretical refinement — an atypical investment.

4. Toward Validity by Design

To retain clarity of evidence for substantive research questions (e.g., whether self-efficacy influences users' information disclosure behavior), it is necessary to gather evidence about measurement validity and the substantive relationship separately: first measurement validation data, then data on the construct's effect of interest [cf. 14]. Since neither advanced statistical methods nor large samples can rectify invalid measures [15], we propose fundamentally changing current HCI research practices by clearly separating these two evidence bases.

The concept of 'validity by design' proposes that measurement rigor must be front-loaded into the overall process of HCI research. Rather than treating scale development as an afterthought or making quick adaptations, the research community would proactively establish evidence for construct validity at the inception of a research program targeting a specific construct. This might involve conducting theoretical construct discussion and empirical evidence supporting the nomological net [12, cf.] before any substantive experiments and surveys begin. This approach clearly separates measurement validation from substantive hypothesis evaluation so that scales undergo psychometric evaluation before they are used to test hypothesized claims. By adopting such an approach, HCI can align itself with best practices, ensuring that its empirical findings rest on a stable measurement foundation rather than an ad-hoc collection of scales.

To practically realize the principle of validity by design, we propose a collaborative, consortium-based approach within the HCI research community. The consortium's first task entails identifying core psychological constructs widely studied in the field — constructs around which researchers can collectively develop consensus on their operational definitions. At this

stage, meta-scientific insights from ongoing jingle-jangle discussions around redundant constructs and high-level parsimonious construct clusters [19, 20] are particularly valuable for narrowing the scope of initial validation efforts. Identifying precisely which constructs warrant priority in these initial validation efforts remains an open question for now. Once identified, the measures of these constructs will undergo substantive development and systematic validation [see three phase process outlined by, 4]. The role of LLM-assisted scale development and the synthetic evaluation of psychometric quality criteria should be discussed and explored as a way to potentially reduce resource intensity [21]. So rather than each research group individually creating or modifying scales, the community would collaboratively within the consortium develop standardized measures that are structurally assessed across multiple contexts and samples [cf. 22]. This also prominently involves external mapping and testing out constructs' nomological networks [see, 5, 8]. We would like to emphasis the incremental nature of validation, in which evidence is collected cumulatively, and hence, can greatly benefit from joint resource allocation within the HCI community.

We suggest a threefold approach to positively impact the adoption rate of validated measures. First, after these initial validation efforts measures will be shared on an open-access platform, which shall serve as an infrastructural tool that helps streamline researchers workflows. By providing transparent and structured instructions for scale implementation and scoring, the platform can enhance the ease of integration into future research. Equally important, it should enable straightforward adoption through ready-to-use resources that have, e.g., seamless import functionalities for items and are compatible with widely used survey software. Second, the consortium must strategically involve influential thought leaders — such as established and upcoming HCI researchers, editors, and committee chairs. These thought leaders shall pledge to using the validated measures themselves, thereby setting standards through their example. Adoption could then triple-down via their top-tier publications and review work, effectively creating community-wide expectations around measurement validity. Finally, the consortium could provide targeted funding and publication incentives explicitly linked to the implementation of validated measurement instruments. For instance, dedicated submission tracks at HCI venues focusing explicitly on validation studies and their applications in various research domains could significantly enhance the visibility of validated scales. Researchers' motivation to engage with the consortium and adopt validated measures would likely increase due to heightened expectations regarding publication opportunities. Additionally, depending on the consortium's funding structure, adoption could be further incentivized by providing financial support (e.g., participant compensation) specifically for research projects employing validated measures.

These recommendations help researchers establish measurement validity from the outset of their projects — facilitating the identification of publishable topics, providing easy access and high salience to state-of-the-art operationalizations of psychological constructs in the literature, offering straightforward retrieval and implementation of validated scales, and potentially even funding related data collection efforts. Such proactive consortium-based support ensures validity is built into research from its inception, bordering on default, and fundamentally reshaping current measurement practices.

Overall, the consortium's aim would be to foster a culture of rigorous, cumulative knowledge-building, thereby strengthening the broader field of HCI. Given that the understanding of constructs is inherently incomplete, the consortium would also serve as a structured plat-

form for ongoing discussions and constructive disagreements about construct definitions and operationalizations, thereby continually refining the field's measurement standards.

5. Conclusion

Just as meta-science in other fields has shown how biases in data collection, publication, or analysis can distort our understanding of phenomena, flawed measurement tools in HCI risk systematically skewing what we think we know about users and their behavior. When HCI scales are developed or adapted without rigorous (re)validation, they can create measurement illusions, whereby researchers and practitioners collectively believe they are tracking an essential user-centered construct (like trust, self-efficacy, or aggression) but are potentially measuring something else entirely. At scale, this invalidity can lead to misguided theory development, product designs, and policy decisions, spawning a cascade of real-world effects that meta-scientists struggle to identify because the distortion began at the foundational level of measurement. We therefore advocate for implementing a 'validity by design' approach in HCI research, proactively integrating systematic validation efforts into the earliest stages of the research process. By forming a collaborative consortium, the HCI community can collectively define core constructs, rigorously validate standardized measurement tools, and openly share validated scales through dedicated infrastructural platforms. Such a consortium can leverage community-wide norms and targeted incentives, facilitating the adoption of validated measures from the outset of research projects.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT-4o and DeepL in order to perform grammar, translation, and spelling checks. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. P. Shonkoff, D. A. Phillips (Eds.), From neurons to neighborhoods: The science of early childhood development, National Academy Press, Washington, DC, USA, 2000. doi:10.17226/9824.

[2] N. Borgert, L. Jansen, I. Böse, J. Friedauer, M. A. Sasse, M. Elson, Self-efficacy and security behavior: Results from a systematic review of research methods, in: F. Mueller, P. Kyburz, J. R. Williamson, C. Sas, M. L. Wilson, P. T. Dugas, I. Shklovski (Eds.), Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24, Association for Computing Machinery, New York, NY, USA, 2024. doi:10.1145/3613904.3642432.

[3] M. Elson, M. R. Mohseni, J. Breuer, M. Scharkow, T. Quandt, Press CRTT to measure aggressive behavior: The unstandardized use of the competitive reaction time task in aggression research, Psychological Assessment 26 (2014) 419–432. doi:10.1037/a0035569.

[4] J. K. Flake, J. Pek, E. Hehman, Construct validation in social and personality research, Social Psychological and Personality Science 8 (2017) 370–378. doi:10.1177/1948550617693063.

[5] L. J. Cronbach, P. E. Meehl, Construct validity in psychological tests, Psychological Bulletin 52 (1955) 281–302.
[6] S. O. Lilienfeld, S. J. Lynn, L. L. Namy, Psychology: From inquiry to understanding, 4 ed., Pearson Education, Boston, MA, USA, 2017.
[7] J. K. Flake, I. J. Davidson, O. Wong, J. Pek, Construct validity and the validity of replication studies: A systematic review, The American Psychologist 77 (2022) 576–588. doi:10.1037/amp0001006.
[8] N. Borgert, O. D. Reithmaier, L. Jansen, L. Hillemann, I. Hussey, M. Elson, Home is where the smart is: Development and validation of the cybersecurity self-efficacy in smart homes (CySESH) scale, in: A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, M. L. Wilson (Eds.), Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, Association for Computing Machinery, New York, NY, USA, 2023. doi:10.1145/3544548.3580860.
[9] A. Bandura, Social foundations of thought and action: A social cognitive theory, Prentice-Hall Series in Social Learning Theory, Prentice-Hall, Englewood Cliffs, NJ, USA, 1986.
[10] A. Bandura, On the functional properties of perceived self-efficacy revisited, Journal of Management 38 (2012) 9–44. doi:10.1177/0149206311410606.
[11] M. Elson, Examining psychological science through systematic meta-method analysis: A call for research, Advances in Methods and Practices in Psychological Science 2 (2019) 350–363. doi:10.1177/2515245919863296.
[12] W. R. Shadish, T. D. Cook, D. T. Campbell, Experimental and quasi-experimental designs for generalized causal inference, Wadsworth Cengage Learning, Belmont, CA, USA, 2002.
[13] S. O. Lilienfeld, A. N. Strother, Psychological measurement and the replication crisis: Four sacred cows, Canadian Psychology/Psychologie Canadienne 61 (2020) 281–288. doi:10.1037/cap0000236.
[14] D. T. Campbell, D. W. Fiske, Convergent and discriminant validation by the multitrait-multimethod matrix, Psychological Bulletin 56 (1959) 81–105. doi:10.1037/h0046016.
[15] J. K. Flake, E. I. Fried, Measurement schmeasurement: Questionable measurement practices and how to avoid them, Advances in Methods and Practices in Psychological Science 3 (2020) 456–465. doi:10.1177/2515245920952393.
[16] I. Hussey, T. Alsalti, F. Bosco, M. Elson, R. Arslan, An aberrant abundance of Cronbach's alpha values at .70, Advances in Methods and Practices in Psychological Science 8 (2025). doi:10.1177/25152459241287123.
[17] M. Elson, I. Hussey, T. Alsalti, R. C. Arslan, Psychological measures aren't toothbrushes, Communications Psychology 1 (2023) 25. doi:10.1038/s44271-023-00026-9.
[18] M. Wischnewski, N. Krämer, E. Müller, Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions, in: A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, M. L. Wilson (Eds.), Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, ACM, New York, NY, USA, 2023. doi:10.1145/3544548.3581197.
[19] G. Hodson, Construct jangle or construct mangle? Thinking straight about (nonredundant) psychological constructs, Journal of Theoretical Social Psychology 592 (2021) 576–590. doi:10.1002/jts5.120.
[20] D. U. Wulff, R. Mata, Semantic embeddings reveal and address taxonomic incommen-

surability in psychological measurement, Nature Human Behaviour 9 (2025) 944––954. doi:10.1038/s41562-024-02089-y.

[21] B. E. Hommel, R. C. Arslan, Language models accurately infer correlations between psychological items and scales from text alone: preprint (2025). doi:10.31234/osf.io/kjuce_v3.

[22] I. Hussey, S. Hughes, Hidden invalidity among 15 commonly used measures in social and personality psychology, Advances in Methods and Practices in Psychological Science 3 (2020) 166–184. doi:10.1177/2515245919882903.