Context is Key for Reproducibility of Empirical Studies in Human-Computer Interaction

Jean Vanderdonckt
Université catholique de Louvain
LouRIM
Louvain-la-Neuve, Belgium
jean.vanderdonckt@uclouvain.be

Radu-Daniel Vatavu
MintViz Lab, MANSiD Research Center
Ştefan cel Mare University of Suceava
Suceava, Romania
radu.vatavu@usm.ro

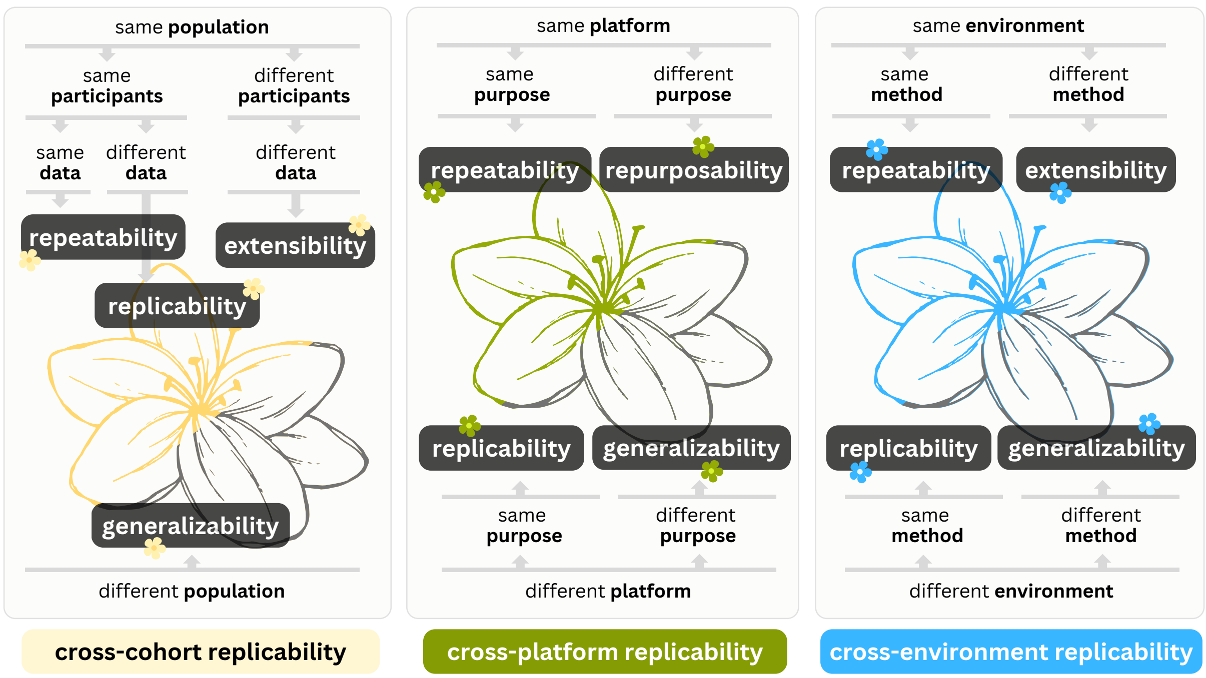

Figure 1: The Ampli❀tum framework for contextual reproducibility in Human-Computer Interaction research.

Abstract

Although most reproducibility frameworks are generic enough to apply across a wide range of scientific domains, they may fail to capture the specific dimensions relevant to particular fields, such as Human-Computer Interaction, where interactive prototypes must be evaluated in the context of use involving users, devices, and environments. To address this aspect, we introduce Ampli❀tum, a contextual reproducibility framework that extends general replication types to cover interactive systems with (1) replication along various contextual dimensions, (2) extension and generalization of previous studies using different populations, and (3) repurposing studies with goals different from those of the original research. To demonstrate the applicability and utility of our framework, we report case studies in gesture-based input. In our first example, we classify existing gesture elicitation studies using our revised terminology. In the second case, we replicate a previous user study under different contextual conditions. Finally, we propose recommendations for the implementation of replicability in HCI.

ACM REP '25, July 29–31, 2025, Vancouver, BC, Canada

ACM ISBN 979-8-4007-1958-5/2025/07
https://doi.org/10.1145/3736731.3746142

CCS Concepts

• **Human-centered computing → Gestural input**; **User interface design**; *Participatory design.*

Keywords

Reproducibility, Reusability, Repurposability, Replicability, Usability, Gesture elicitation studies

ACM Reference Format:

Jean Vanderdonckt and Radu-Daniel Vatavu. 2025. Context is Key for Reproducibility of Empirical Studies in Human-Computer Interaction. In *ACM Conference on Reproducibility and Replicability (ACM REP '25), July 29–31, 2025, Vancouver, BC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3736731.3746142

1 Introduction

The reproducibility of research artifacts, ranging from software and datasets to scripts, raw data, and hardware designs, has become a cornerstone of scientific and technological progress. While existing frameworks, such as those from the Association of Computing Machinery (ACM) [1,2], focus on ensuring that artifacts are reproducible within precisely defined setups, the diverse contexts of use in which these artifacts are used present unaddressed challenges. For example, the ACM review and badging system [2] employs two dimensions—*team* and *experimental setup*—to define *repeatability* (same team and same experimental setup), *reproducibility* (different team, same experimental setup), and *replicability* (different team, different experimental setup) of artifacts, where an artifact is any "digital object that was either created by the authors to be used as part of the study or generated by the experiment itself" [2].

While this classification is broadly applicable in science and engineering, it does not completely capture the fine granularity of experiments and studies conducted in specific fields, such as Human-Computer Interaction (HCI). In HCI, artifacts, such as interactive applications and their user interfaces, are typically evaluated in relation to the context in which they will be used. The *context of use* subsumes three dimensions [11,17,18]: the *end user* (e.g., a user profile) and the corresponding interactive tasks (e.g., through a task model), the *device/platform* used to carry out these tasks (e.g., smartphone, laptop, electronic ring, etc.), and the *environment* in which the tasks are performed (e.g., the physical setting and the organizational and psycho-social aspects of the environment). Traditional approaches to reproducibility often focus on the ability to replicate results under identical experimental conditions, implicitly assuming the same context of use. However, interactive systems and the complexities of modern scientific research require a broader, context-sensitive approach to reproducibility.

Contextual evaluation means that an artifact is evaluated in an explicitly defined context of use. Unfortunately, this context is not always clearly articulated in experiments or research papers, which obscures the conditions in which the artifact might be applicable or reproducible. This is a significant drawback since variations in contextual dimensions, such as user experience, computational platforms, operating environments, and application domains, can impact reproducibility, making it imperative to develop a more flexible framework representative of these dimensions. Furthermore, there is no formalization of how replications could be implemented for the specifics of contextual evaluation or how existing findings could be generalized across diverse contexts of use [11].

A framework that extends artifact reproducibility across contexts of use, addressing the complex interplay of technical, human, and environmental factors, would benefit a broad range of stakeholders, including the potential for reproducibility in interdisciplinary research involving human factors. This framework should ensure the credibility and longevity of research and foster innovation by enabling broader adoption and adaptation of artifacts. To address this need, we present in this work Ampli❀tum, an extended framework for contextual reproducibility in HCI. Borrowed from Latin, we use the term "ampliatum" to denote the extension or broadening of existing replicability concepts to cover the specific challenges of replication in HCI. The term also appears in the name of orchid species, such as Oncidium ampliatum, a yellow-flowered orchid with distinctive pseudobulbs, which we use in this work for visual representations of our framework; see Figure 1. To this end, we make the following contributions:

(1) We introduce Ampli❀tum, a framework for contextual reproducibility of artifacts in the field of HCI, implementing specific reproducibility types involving users, platforms, and environments in conjunction with the collected data, the research method, and the researchers' goals; see Figure 1.
(2) We define new terminology to refer to specific types of replication addressing users (*cross-cohort replicability*), platforms (*cross-platform replicability*), environments (*cross-environment replicability*), and contexts of use (*cross-context replicability*), which refine existing replication concepts according to key elements of the context of use [11].
(3) To showcase the applicability and utility of Ampli❀tum and its corresponding terminology, we discuss replications in gesture elicitation studies [59], a specific type of participatory design in which end users propose gesture commands they would like to use for interactive systems. We also conduct new replications in different contexts of use of previous gesture elicitation studies involving electronic rings [21,22].

2 Related Work

2.1 Reproducibility Outside Computer Science

Reproducibility is the foundation for cumulative scientific knowledge and progress, ensuring that research findings are consistent, reliable, verifiable, and constitutes a fundamental prerequisite to establish scientific validity [27,43]. Reproducibility requires clear documentation of methods, experimental designs, and procedural details, along with the availability of relevant resources. The absence of publicly accessible data presents a real barrier to reproducibility, repeatability, and verifiability, as the data collection process can be arduous and resource intensive [41].

In recent years, the issue of reproducibility has attracted increasing attention. Although many studies adhere to rigorous methodologies and maintain transparency in their procedures, reproducibility extends beyond replicating an experimental setup. Instead, it requires ensuring that the research findings remain valid and reliable in different contexts of use [8] involving variations in the team conducting the research, the platforms employed, and the environmental conditions under which experiments are administered [11]. The term "reproducibility crisis" [45] describes the difficulty in replicating scientific studies, raising concerns in multiple disciplines.

A variety of conceptual frameworks have been proposed to improve reproducibility in disciplines outside of computer science. Gómez et al. [25] identified twenty-seven distinct frameworks for classifying reproducibility. Dreber and Johannesson [19] developed a typology for assessing reproducibility and replicability in economics, categorizing studies into five types (i.e., computational reproducibility, recreate reproducibility, robustness reproducibility, direct applicability, and conceptual applicability) and proposing specific indicators for evaluation (e.g., comparisons across different populations). In health and social sciences, Vissoci et al. [56] presented a framework for reproducible research, emphasizing the integration of open-source technologies in data analysis and dissemination. Within medical research, Pokutnaya et al. [40] proposed a framework comprising twenty-two elements in six categories that aim to improve the transparency and reproducibility of computational models of infectious diseases. Choi [12] developed a reproducible framework to ensure transparency and quality in nursing research to promote evidence-based practice.

2.2 Reproducibility Inside Computer Science

Several frameworks have been proposed to characterize the reproducibility of experiments, such as formalizing reproducibility in empirical software engineering [16]. For example, Anchundia and Fonseca C. [5] identified tools for maximizing reproducibility, noting differences between human-oriented and technology-oriented experiments. Mendonça et al. [35] proposed FIRE, the Framework for Improving the Replication of Experiments, to address knowledge sharing issues in replications. González-Barahona and Robles [26] explored elements characterizing reproducibility in studies based on development repository data, emphasizing the need for a systematic approach to reproducibility. Cordeiro [15] proposed a framework to promote reproducibility in software engineering experiments focusing on provenance, preservation, and curation of research artifacts. Kedron et al. [31] advocated for an approach to geographical analyzes that goes beyond computational reproducibility, emphasizing the importance of evaluating the validity of the conclusion and its evidence. Also, Mondelli et al. [36] proposed a conceptual framework ensuring the reproducibility of computational experiments while adhering to the FAIR (Findable, Accessible, Interoperable, Reusable) principles [7]. Patil et al. [38] introduced a visual notation to indicate how reproducibility is addressed. Peng [39] argued that scientific articles are not replicable unless their artifacts are made available by releasing data and source code [20].

2.3 Reproducibility in HCI

The HCI research community has also acknowledged the importance of reproducibility, namely via RepliCHI [57]. However, the replication rate, evaluated in 2014, was exceptionally low, about 3% [29]. To advance the field, the HCI community must further embrace reproducibility, particularly in cases where findings challenge prior results [28]. Noteworthy, the term "replication" has been widely adopted within HCI, whereas "reproducibility" has remained prevalent in other scientific disciplines [27]. Unfortunately, this distinction in terminology has contributed to conceptual ambiguity, a challenge that also extends across multiple scientific fields.

2.4 Gesture Elicitation Studies

To illustrate Ampli❀tum, we present in Sections 5 and 6 two case studies involving gesture elicitation. In the following, we briefly review the corresponding methodology. A gesture elicitation study (GES) [58,59] presents participants in an experiment with the effects following a command in an interactive system, and elicits gestures that trigger those effects. For example, Zaiţi et al. [60] elicited mid-air gestures to control various functions of a smart TV, such as "turn the TV on and off" and "go to the next and previous channels." These system functions are called referents [59], for which the user-proposed input may take the form of any symbol or action relevant to the interactive system, such as gesture commands, key presses, eye gaze input, or voice commands. Subsequently, the practitioner analyzes the elicited proposals to understand the level of consensus among users. RepliGES [24] is a conceptual space for conducting replications of gesture elicitation studies, which we use as a foundation for our exploration; see next.

3 Foundations

We build on previous classifications of reproducibility in scientific research [27,39,50], ACM's definitions of artifact review and badging [1,2], and the replication types introduced by Gheran et al. [24] in the specific context of conducting end-user elicitation studies in HCI (see Figure 2). Next, we briefly summarize these concepts.

3.1 Artifacts

ACM [2] defines an artifact as "a digital object that was created by the authors to be used as part of the study or generated by the experiment itself." In general computer science, artifacts may include software, electronic components, system models and architectures, or data collected with or generated by these systems. In HCI, artifacts often take the form of interactive prototypes, user interfaces, experimental designs, or the data collected from participants in a user study. Following [24], we differentiate between reviewing the same data (verifying the analysis), collecting new data from the same participants (repeating the study) and collecting new data from new participants in relation to specific artifacts.

3.2 Research Team and Experimental Setup

The method of analysis, proposed by Tsang and Kwan [50], specifies possible replication types, of which ACM [2] implements variations in team and setup. However, we note a key omission in the ACM's [2] team dimension, where the configuration "same team, different setup" is not covered, although relevant when the same team wishes to conduct a variation of the original study to expand on previous findings in relation to a variation in the context of use.

3.3 Types of Reproducibility

Goodman et al. [27] highlighted three types of reproducibility: (i) *method reproducibility* strives to make the data and experimental setup fully accessible so that the experiment can be repeated; (ii) if the experiment can be repeated with the same results, *results reproducibility* is accomplished; and (iii) with *inferential reproducibility*, the same conclusion is obtained by another inference method, such as a different statistical test. Besides these types, we argue in this paper that *contextual reproducibility* is the key to the field of HCI, where interactive prototypes, user performance, and user experience are studied in context. In terms of both method and results reproducibility, contextual reproducibility strives for generalization, extension, and repurposing of artifacts for new discoveries in HCI.

3.4 Types of Replications

Tsang and Kwan [50] classified replications on two dimensions: *sources of data* (same data set, same population, and different populations) and *changes in the method/procedure* (same measurement and analysis, different measurements and/or analysis). Six types were defined: checking of analysis (which corresponds to ACM's [2]

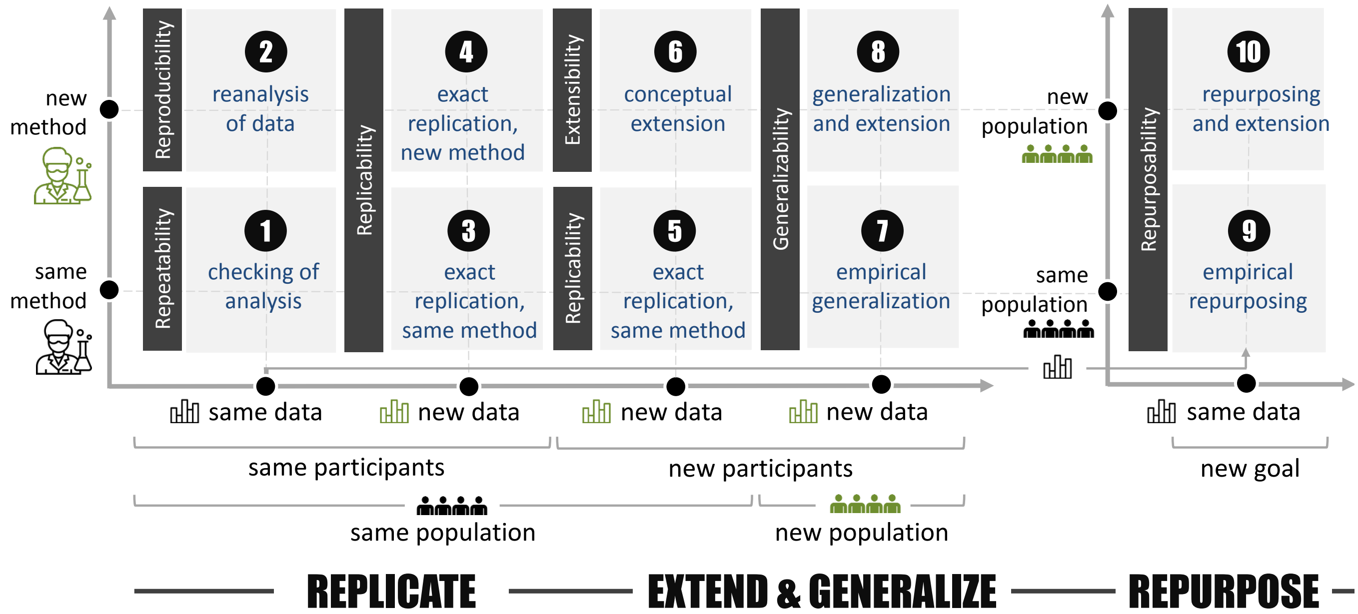

Figure 2: The RepliGES space, reproduced from Gheran et al. [24], originally introduced for end-user gesture elicitation studies. From this space, we adopt the concepts of extension, generalization, and repurposing to describe reproducibility in context in Ampli❀tum. Note that two key elements of the context of use, *platform* and *environment*, are not present in RepliGES.

repeatability), reanalysis of data (different procedures are used on the same data), exact replication (the same study conducted with new participants from the same population), conceptual extension (new procedure applied to new participants from the same population), empirical generalization (the same procedure used as in the original study but on a new population), and generalization and extension (new procedure and new population). The team involved in the replication is not considered important, but the population is. Since generalizations of the findings of one study to other populations can be insightful in consolidating HCI knowledge about how users engage with interactive systems, we rely on the *population* dimension next to those of *data* and *participants* by considering the scenario in which new data are collected from participants sampled from a population different from the one in the original study.

In the field of HCI, Hornbæk et al. [29] distinguished among three types of replication: a *strict replication*, which reuses the same variables from the original study, a *partial replication*, which introduces controlled variations to investigate how the original results hold in a different setup, and a *conceptual replication*, which applies other measurement methods. Moreover, the following specific replication types were proposed in the RepliGES framework introduced by Gheran et al. [24] (Figure 2):

- *Repeatability.* The data from the original study are reanalyzed with the same method to verify if the same results are obtained. This replication type connects to Tsang and Kwan [50] "checking of analysis" and ACM's [2] "repeatability".
- *Reproducibility.* The data from the original study is revisited with a new method, while maintaining the research question. The study may be conducted by an independent team, as in ACM's [2] definition of "reproducibility," or by the original researchers, i.e., what Tsang and Kwan refer to as a "reanalysis of data" [50].
- *Replicability.* The research team collects new data, either from the same participants of the original study or new participants to verify if the original results can be confirmed.
- *Extensibility.* The research team employs a new method on the data collected from new participants, i.e., they perform a "conceptual extension" [50] of the original study.
- *Generalizability.* The research team runs the study by employing the original method, but with a sample of participants from another population to check whether the original findings generalize to the new population.
- *Repurposability.* The data are used for a new goal to discover something new about the same or a new population.

It is important to note that Gheran et al. [24] applied these replication types across users, populations, and methods within the specific application domain of gesture elicitation studies, but did not address platforms or environments. In the next section, we extend their framework toward contextual reproducibility in HCI.

4 Ampli❀tum, a Framework for Contextual Reproducibility in HCI

Building on the prior work reviewed in the previous sections, we highlight the need for specific types of replication and corresponding terminology to better support both the development of interactive prototypes and the reproducibility of empirical findings in interactive systems. In Subsection 4.1, we summarize the key elements that specify the context of use. In Subsection 4.2, we introduce extensions to existing frameworks to address the specifics of HCI replications in relation to context, instantiated in Ampli❀tum.

Table 1: Terminology for contextual reproducibility in HCI studies and interactive systems.

Users	Platforms	Environment	Interpretation / Type of study	Terminology
=	=	=	All conditions are the same	Repeatability
≠	=	=	Different users	Cross-cohort replicability
=	≠	=	Different platforms	Cross-platform replicability
=	=	≠	Different environments	Cross-environment replicability
=	≠	≠	Different platforms and environments	Cross-context replicability (variant)
≠	≠	=	Different users engaging with different platforms	Cross-context replicability (variant)
≠	=	≠	Different users in different environments	Cross-context replicability (variant)
≠	≠	≠	Different users, platforms, and environments	Cross-context replicability (variant)

Legend: = denotes the same condition and ≠ a different condition in terms of the elements of the context of use.

4.1 Elements of the Context of Use

The three key pillars of the context of use [11,17,18]—users and their tasks, devices and corresponding platforms, and environments—provide a holistic approach for contextual reproducibility.

4.1.1 Users and tasks. Reproducibility is defined not just by identical conditions, but also by ensuring that both implementation and results are consistent across a range of typical users and tasks. However, different users may interpret instructions, data, and tasks in varying ways and, consequently, their background, expertise, and goals can significantly influence reproducibility outcomes. To address this aspect, protocols and documentation are needed to explicitly present user roles (e.g., their job), tasks (e.g., duties assigned to jobs), and goals (e.g., high-level mission statement). Adaptation of tasks and procedures should be accounted for meaningful contextual replication during the scientific investigation.

4.1.2 Devices and platforms. Reproducibility encompasses achieving consistent implementation and identical or equivalent results on diverse devices and computational platforms (e.g., smartphones, desktop computers, smart wearables such as electronic rings, tabletops, cloud services, and supercomputers). Differences in hardware, operating systems, and software configurations across these devices and platforms can lead to variations in user performance, ultimately affecting replication outcomes. Reproducibility frameworks should therefore account for this dimension through conceptual extension and refinement, ensuring that results hold across different devices and platforms, rather than just replicating the computational environment. The corresponding methods and tools should also be compatible with a wide range of devices and platforms, e.g., standardization of software packages, portability of data formats, and implementation of computational workflows represent essential technical aspects in the engineering of interactive systems to ensure that results remain consistent during replications.

4.1.3 Environments. The physical, social, and organizational environment in which HCI research is conducted also impact reproducibility outcomes. For example, factors like laboratory conditions, network bandwidth, access to online resources, and collaborative frameworks involving multi-user [46], multi-device [10], and multi-component [52] interactions can impact the execution and results of scientific work. Environments with unfamiliar, unconventional, or extreme characteristics [51] further increase the complexity caused by these factors. Consequently, reproducibility frameworks should explicitly account for environmental characteristics.

4.2 Types of Contextual Replications

To represent the dimensions of the context of use in relation to replication challenges, we use the flower metaphor in Figure 1 to distinguish among the *user* (i.e., same or different participants), the *population* (i.e., same or different population), and the *data* (i.e., same or different). We apply the same approach to the *platform* and *environment* elements of the context of use in relation to the research method and the researchers' goal. For example, reconducting a study with the same platform and the same goal leads to repeatability, while changing both the environment and the method leads to generalizability. Based on these concepts, we define the following types of contextual replications in HCI; see Table 1.

4.2.1 Cross-cohort replicability. This involves reproducing a study or experimental procedure using a new group of participants who differ meaningfully from those in the original study. Differences may cover variations in population sampling, demographic factors (e.g., age, gender, profile), geographic characteristics (e.g., ethnicity, country, continent), and social, cultural, or temporal aspects. This approach tests the generalizability and robustness of findings in diverse populations. For example, a researcher might wish to validate the scalability of a machine learning model by running a cross-cohort replication based on a dataset from a user demographic different from the original training set. A study conducted with the same participants as the original is described as co-cohorted whereas one with different participants is called re-cohorted.

4.2.2 Cross-platform replicability. In this case, reproducing a study or experimental procedure (e.g., a system evaluation, a user interaction experiment) involves a computing platform that is technically or functionally different from the one used in the original study. This may involve differences in software (e.g., operating systems, network protocols), hardware (e.g., desktop vs. mobile vs. wearable devices), or interaction capabilities (e.g., voice vs. gesture input as part of natural interaction). This approach tests the generalizability and robustness of the findings across diverse technological spaces [30]. For example, a researcher might wish to evaluate the generalizability of an interaction technique by running a cross-platform replication on a mobile device, in contrast to the original desktop-based implementation. Another example is computational cross-platform applicability, which is abstract enough to encompass both software and hardware changes. A study carried out with the same platforms as the original is described as co-instrumented, whereas one with a different platform is called re-instrumented.

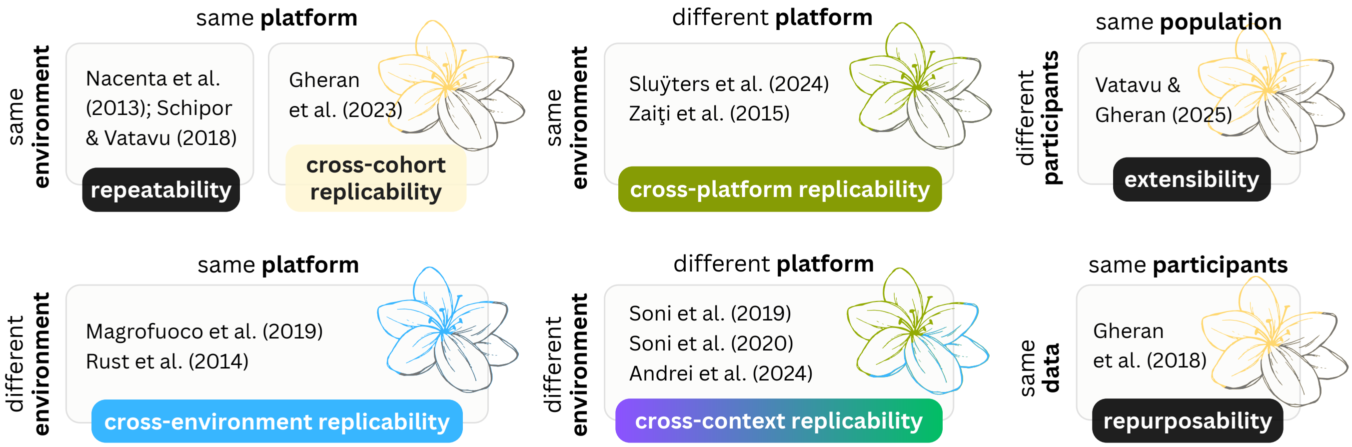

Figure 3: Examples of published studies representing various replication types according to the categories of Ampli❀tum.

4.2.3 Cross-environment replicability. This involves reproducing a study or experimental procedure in a physical environment that differs significantly from the setting used in the original study. Differences may cover aspects of mobility (e.g., moving the study from a laboratory to a real-world deployment), geo-relocation (e.g., from one geographic location to another), or between different infrastructural, cultural, organizational, and institutional environments (e.g., from the headquarters to a local agency). This approach tests the generalizability, environmental robustness, and the ecological validity of the original findings. For example, a researcher wishes to examine the ecological validity of previous results by carrying out a cross-environmental replication in a public co-working space (*in situ*), contrasting with the original lab-based environment (*ex situ*). A study carried out in the same environment as the original is described as co-located, whereas one involving a different environment is called re-located.

4.2.4 Cross-context replicability. This type of replication covers any combination of the above, involving a new context of use that differs from that of the original research. Any form of cross-type replicability constitutes a particular case of cross-context replicability; see Table 1. A study conducted within the same context of use as the original is described as co-contextualized whereas one involving a different context is called re-contextualized.

5 Situating Existing Replications with Ampli❀tum

In this section, we illustrate the use of Ampli❀tum to position existing replications in HCI, for which we adopt the use case of gesture elicitation studies. In these studies, users propose gesture input they would like to use to control interactive systems, such as multi-touch gestures for interactive tabletops [59].

Figure 3 shows examples of replications conducted in previously published studies across cohorts, platforms, and environments. For example, Nacenta et al. [37] re-invited the same participants in the same environment after twenty-four hours to investigate whether they still remembered the gestures proposed a day ago, also maintaining the same platform; see the top-left part in Figure 3. According to our terminology, this involves conducting a repeatability of the original study. Schipor and Vatavu [44] is another example of repeatability by involving the same participants, whereas Gheran et al. [21] involved the same environment, same platform, but new participants, representing a cross-cohort replication.

Sluyters et al. [47] performed replication across a different platform, where they re-computed the agreement rate for gestures proposed in the same experimental room, but using a different platform to ensure the correctness of the calculations; see Figure 3, top-middle. Zaiţi et al. [60] used a different platform, represented by a Leap Motion controller, to elicit mid-air hand gestures for a smart TV setup in the same environment as a previous study [53]. By changing the environment and maintaining the same platform, Magrofuoco et al. [32] conducted a gesture identification study to refine and check the results of a previous elicitation study; different participants were involved, but the experimental setup remained consistent. Similarly, Rust et al. [42] adapted Wobbrock et al.'s [59] study on tabletop interaction by eliciting gestures from children, carried out in a different environment but with a similar platform; see Figure 3, bottom left. When both the environment and platform are different, cross-context replicability is achieved. For example, Soni et al. [48] compared gestures proposed by children and adults on spherical displays to gestures proposed on flat tabletops [49], while Andrei et al. [6] elicited on-chair gestures involving a different platform (armchair, office chair, and stool) in a laboratory environment, replicating a previous study where wheelchair users were elicited about their preferences for on-wheelchair input [9].

Within the same population, collecting data from different participants can lead to cross-cohort replicability. For example, Gheran et al. [21] replicated a gesture elicitation study involving smart rings to understand whether the same gesture set would merge again with new participants, five years after the original study [22]. Within the same environment, Vatavu and Gheran [54] analyzed whether on-body gestures proposed by different participants on different regions on the arm possess the same articulation characteristics. Lastly, Gheran et al. [23] reutilized existing data [22] to report new discoveries about the articulation of bimanual gestures detected with electronic rings worn on both hands; see Figure 3, right for the extensibility and replicability categories.

Table 2: Comparison of several contextual replications of a gesture elicitation study [22].

Study	n	Population	Method	Team	Goal	Platform	Environment
Original study [22]	24	General	Agreement rate, thinking time, goodness of fit	3	Consensus set	Ring Zero	Laboratory
Replication [21]	=	=	=	=	=	=	=
Repurposing data [23]	=	=	Allen's Interval Algebra ≠	=	Classification	=	=
Study 1 (Section 6)	30 ≠	=	=	5 ≠	Generalization	=	Office
Study 2 (Section 6)	30 ≠	=	=	6 ≠	Generalization	=	Library
Study 3 (Section 6)	30 ≠	=	=	6 ≠	Generalization	=	Office
Study 4 (Section 6)	30 ≠	=	=	6 ≠	Generalization	=	Classroom

Legend: = denotes the same condition, ≠ denotes a different condition, e.g., the original team was composed of three researchers, while the generalizations involved teams of at least five researchers.

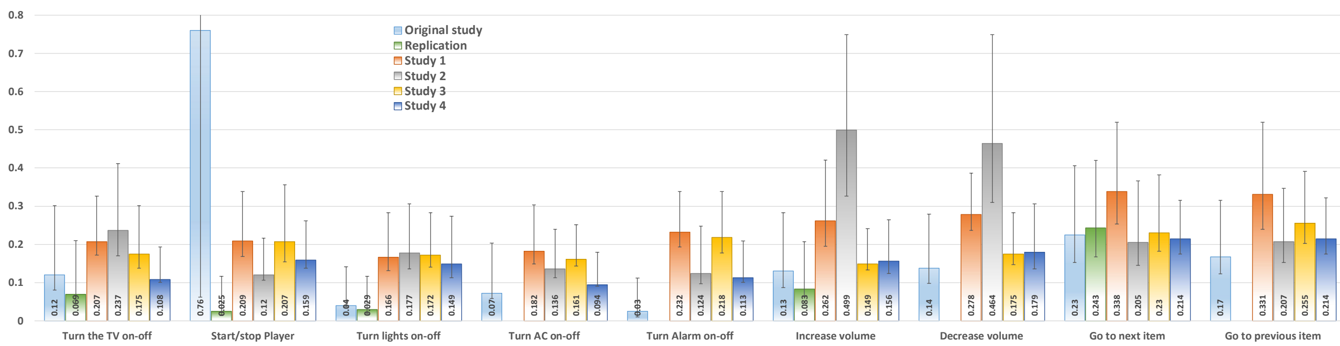

Figure 4: Agreement rates from various replications of an original study [22]. Error bars show 95% CIs.

6 Conducting New Replications with Ampli❀tum

The results of a GES are typically represented by a consensus set, which contains the gestures that occur the most frequently among the participants' proposals, validated using agreement measures, primarily using agreement scores [59] and agreement and co-agreement rates [55]. Several tools are available to help practitioners implement GESs and analyze the elicited gestures [3,4,33,55].

To examine the effects of different types of replication carried out in different contexts, we start with a study for which the data are publicly available—[22], which was replicated in [21] for repeatability and extended in [23] for repurposability. We designed four replications, hereafter named Study 1 to Study 4, each involving a different sample of 30 participants. The participants were instructed to propose gestures for eleven referents. The method was the same as in [22], consisting of calculating the agreement rates, measuring thinking times, and eliciting ratings of the goodness of fit between proposed gestures and the system functions they executed. Each study was conducted by a different team comprising five or six experimenters to preserve the autonomy of administration, reduce the effect of habituation, and avoid the Hawthorne effect [34].

The repurposability of the original data reported in [23] did not change the context of use, since data was already collected. The replication reported in [21] represented a cross-cohort replication since the participants were new, involving the same research team, platform, and environment. Each of our four studies represents both a cross-cohort replication by involving new participants from the same population, and a cross-environment replication by placing the participants in different environments compared to the one in the original study. Therefore, they also represent cross-context replications. The main purpose of running these studies was to understand the generalizability of previous results with the same device, but in different environments: an office, a library, and a classroom setting. Table 2 compares the replication [21], the repurposability study [23], and our four studies addressing generalizability.

Figure 4 shows the agreement rates, computed using AGATe [55], across individual referents. On averaged, we observed .161 agreement in the original study, .090 in the replication, and .272, .226, .246, and .169 in the four new studies, respectively. The replication received the lowest average, the original study and Study 1 are comparable, and Studies 1 to 3 are also similar. Since the six replications do not share the same number of referents, we performed two statistical tests: a first set of t-tests on independent samples involving all referents and a second set on the common referents only. In sum, a very limited number of statistically significant differences exists among studies in terms of agreement rate: between the original study and Study 1 (p=.027) with a small effect size (r=.360) on all referents, between Study 1 and Study 4 ($p < .0001$) with a large effect size (r=.79) on their shared 15 referents, between Study 3 and Study 4 (p=.004) with a medium effect size (r=.630), and between

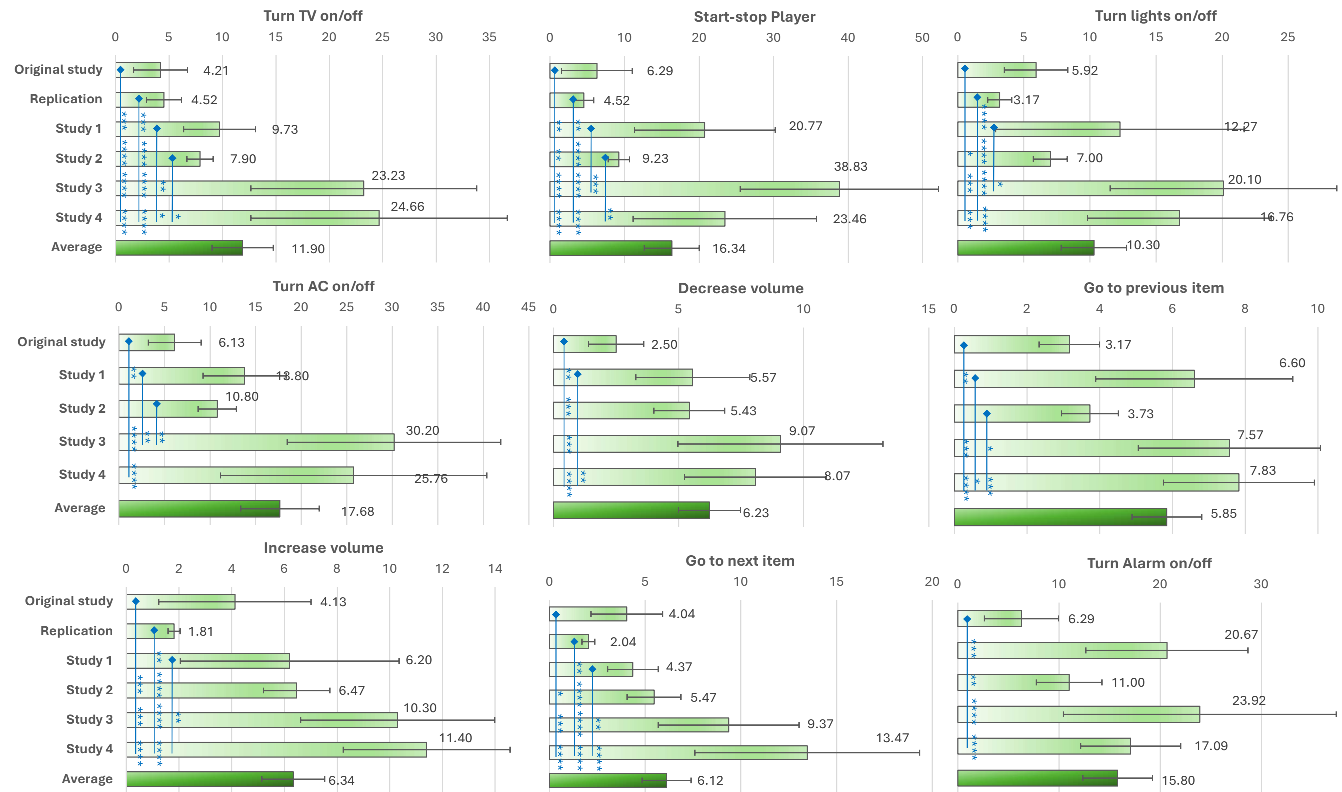

Figure 5: Thinking time measured in replications with respect to the original study. Error bars show 95% CIs.

Study 2 and Study 4 (p=.042<.05), which are represented in Fig. 4. No other t-test was revealed significant, thus suggesting that the six replications did not produce many different agreement rates for their respective parts, apart from a few cases.

The thinking time corresponds to the time, measured in seconds, needed by a participant to propose a gesture for effecting a given referent, originally used in [22]. To understand how thinking time performed across the replications, we computed averages per individual referents; see Figure 5. The shortest time (6.1s) was observed in the original study while the longest (30.2 sec) in Study 3. For example, the "Turn TV on/off" referent received the shortest time (4.2s) in the original study, yet the longest one (24.7s) in Study 4, revealing differences caused by the cross-cohort and cross-environment replications. Kruskall-Wallis tests showed that the average thinking time was statistically different among the replications ($H_{(5)}$=500.84, $p \leq .0001$), and post-hoc Conover-Iman tests [14] revealed that, in most cases, the original study and the replication were different than subsequent generalizations. These differences are explained by variations in population sampling (cross-cohort replicability) and setting (cross-environment replicability); see Table 2. When a significant difference was detected, the effect size, calculated using Cohen's d [13], was small to medium.

7 Discussion

In this section, we offer practical recommendations for implementing contextual reproducibility in HCI. While these take the form of heuristics inspired by our specific application domain, we believe they are broadly applicable to other fields where the context in which artifacts are created, used, and studied plays a critical role.

Defining a study in HCI means controlling possible variations in term of users, platforms, and environments, once the goal of the study, the method, and the target population have been determined. When context [11] is not explicitly specified or controlled, the factors not catered for can easily lead to a combinatorial explosion of contextual possibilities and configurations. However, once the target population has been specified, the number of participants is determined according to objective aspects in generic experimental science, such as statistical power, and subjective ones, such as the experimenter's goal. Results can be verified and validated through repeatability by the same team, and reproducibility by a different team or a different method [1,2]. Beyond reproducibility, new data from new participants, possibly representing a sample from a new population, can lead to extensibility, as in the RepliGES framework [24]. In this case, cross-cohort replications enable researchers to understand how original results hold in situations where the end users of an interactive system present different demographics.

New in Ampli❀tum, cross-platform replication is the first step to diversifying the context of use, e.g., to investigate the potential effects of the specific platform on results about user performance or experience. We believe that cross-platform replications are increasingly relevant in HCI given the evolving landscape of input and output devices, sensors, and operating systems featuring various functionalities. In particular, transitioning across interaction paradigms, from desktop to mobile to wearable computing, needs replication studies to understand the effect of specific platforms on users' capacity to interact effectively and enjoy the corresponding experience. The following step is cross-environment replication, where the potential effects of the setting in which the study is conducted can influence its results. We argue that cross-environment replications become especially important in the context of mixed worlds, featuring diverse combinations of physical and virtual.

When multiple elements of the context of use change, cross-context replication is achieved. However, compared to the previous cases, this comes at the expense of more factors that can vary, making both the experiment design and data analysis more complex. Increasing the number of researchers in the team, splitting replications across teams, and limiting the number of conditions for each factor may prove beneficial to reduce team-induced treatment bias and the risk of observing the Hawthorne effect.

8 Conclusion

We introduced Ampli❀tum, a framework for contextual reproducibility in the field of Human-Computer Interaction to address shortcomings of existing generic replicability frameworks. To showcase our framework, we reported findings from four new replications of an existing study involving user-defined gestures for smart rings. Contextual reproducibility enhances existing reproducibility types with elements specific to the context of use—users, platforms, environments—relevant for validating scientific contributions in HCI. Ampli❀tum is intended for practitioners, who can rely on their body of knowledge and the replication opportunities enabled through contextual reproducibility to gain new insights and validate existing information for interactive systems. Our framework is also intended for HCI researchers to inform new experiments and studies toward new findings about user behavior, performance, and experience with interactive systems in new contexts of use.

Open Science

The datasets used in this paper are available at https://osf.io/5mqx6.

Acknowledgments

Jean Vanderdonckt acknowledges support from the EU EIC Pathfinder Awareness Inside challenge Symbiotik project (1 Oct. 2022-30 Sept. 2026) under Grant no. 101071147.

References

[1] ACM. 2018. Artifact Review and Badging – Version 1.0 (not current). https://www.acm.org/publications/policies/artifact-review-badging
[2] ACM. 2020. Artifact Review and Badging - Current. https://www.acm.org/publications/policies/artifact-review-and-badging-current
[3] Abdullah X. Ali, Meredith Ringel Morris, and Jacob O. Wobbrock. 2018. Crowdsourcing Similarity Judgments for Agreement Analysis in End-User Elicitation Studies. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) *(UIST '18)*. ACM, New York, NY, USA, 177–188. https://doi.org/10.1145/3242587.3242621
[4] Abdullah X. Ali, Meredith Ringel Morris, and Jacob O. Wobbrock. 2019. Crowdlicit: A System for Conducting Distributed End-User Elicitation and Identification Studies. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. ACM, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300485
[5] Carlos E. Anchundia and Efraín R. Fonseca C. 2020. Resources for Reproducibility of Experiments in Empirical Software Engineering: Topics Derived from a Secondary Study. *IEEE Access* 8 (2020), 8992–9004. https://doi.org/10.1109/ACCESS.2020.2964587
[6] Alexandru-Tudor Andrei, Laura-Bianca Bilius, and Radu-Daniel Vatavu. 2024. Take a Seat, Make a Gesture: Charting User Preferences for On-Chair and From-Chair Gesture Input. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, New York, NY, USA, Article 555, 17 pages. https://doi.org/10.1145/3613904.3642028
[7] Michelle Barker, Neil P. Chue Hong, Daniel S. Katz, Anna-Lena Lamprecht, Carlos Martinez-Ortiz, Fotis Psomopoulos, Jennifer Harrow, Leyla Jael Castro, Morane Gruenpeter, Paula Andrea Martinez, and Tom Honeyman. 2022. Introducing the FAIR Principles for Research Software. *Scientific Data* 9, 622 (2022). https://doi.org/10.1145/2833093
[8] Jay J. Van Bavel, Peter Mende-Siedlecki, William J. Brady, and Diego A. Reinero. 2016. Contextual Sensitivity in Scientific Reproducibility. *Proceedings of the National Academy of Sciences* 113, 23 (2016), 6454–6459. https://doi.org/10.1073/pnas.1521897113
[9] Laura-Bianca Bilius, Ovidiu-Ciprian Ungurean, and Radu-Daniel Vatavu. 2023. Understanding Wheelchair Users' Preferences for On-Body, In-Air, and On-Wheelchair Gestures. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, New York, NY, USA, Article 78, 16 pages. https://doi.org/10.1145/3544548.3580929
[10] Frederik Brudy, Christian Holz, Roman Rädle, Chi-Jui Wu, Steven Houben, Clemens Nylandsted Klokmose, and Nicolai Marquardt. 2019. Cross-Device Taxonomy: Survey, Opportunities and Challenges of Interactions Spanning Across Multiple Devices. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 28 pages. https://doi.org/10.1145/3290605.3300792
[11] Gaëlle Calvary, Joëlle Coutaz, David Thevenin, Quentin Limbourg, Laurent Bouillon, and Jean Vanderdonckt. 2003. A Unifying Reference Framework for multi-target user interfaces. *Interacting with Computers* 15, 3 (2003), 289–308. https://doi.org/10.1016/S0953-5438(03)00010-9
[12] Soyoung Choi. 2021. Evidence-Based Research in Nursing Science: A Reproducible Framework. *Western Journal of Nursing Research* 43, 9 (2021), 828–833. https://doi.org/10.1177/0193945920977792
[13] Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge, New York, NY, USA. https://doi.org/10.4324/9780203771587
[14] W.J. Conover and Ronald N. Iman. 1979. *On Multiple-Comparisons Procedures*. Informal Report LA-7677-MS. University of California, Los Alamos Scientific Laboratory, Los Alamos, New Mexico, USA. https://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-07677-MS
[15] André Cordeiro. 2023. Towards a Framework Based on Open Science Practices for Promoting Reproducibility of Software Engineering Controlled Experiments. In *Anais do XXVI Congresso Ibero-Americano em Engenharia de Software*. SBC, Porto Alegre, RS, Brasil, 237–244. https://doi.org/10.5753/cibse.2023.24707
[16] André Felipe R. Cordeiro. 2022. An Open Science-Based Framework for Managing Experimental Data in Software Engineering. In *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering* (Gothenburg, Sweden) *(EASE '22)*. ACM, New York, NY, USA, 342–346. https://doi.org/10.1145/3530019.3535348
[17] Joëlle Coutaz, James L. Crowley, Simon Dobson, and David Garlan. 2005. Context is Key. *Commun. ACM* 48, 3 (2005), 49–53. https://doi.org/10.1145/1047671.1047703
[18] Anind K. Dey. 2001. Understanding and Using Context. *Pers. Ubiquitous Comput.* 5, 1 (2001), 4–7. https://doi.org/10.1007/S007790170019
[19] Anna Dreber and Magnus Johannesson. 2024. A Framework for Evaluating Reproducibility and Replicability in Economics. *Economic Inquiry* (2024), 1–19. https://doi.org/10.1111/ecin.13244
[20] Florian Echtler and Maximilian Häußler. 2018. Open Source, Open Science, and the Replication Crisis in HCI. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18)*. ACM, New York, NY, USA, 1–8. https://doi.org/10.1145/3170427.3188395
[21] Bogdan-Florin Gheran, Radu-Daniel Vatavu, and Jean Vanderdonckt. 2023. New Insights into User-Defined Smart Ring Gestures with Implications for Gesture Elicitation Studies. In *Extended Abstracts of the ACM CHI Conference on Human Factors in Computing Systems, CHI EA 2023, Hamburg, Germany, April 23-28, 2023*. ACM, 216:1–216:8. https://doi.org/10.1145/3544549.3585590
[22] Bogdan-Florin Gheran, Jean Vanderdonckt, and Radu-Daniel Vatavu. 2018. Gestures for Smart Rings: Empirical Results, Insights, and Design Implications. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, New York, NY, USA, 623–635. https://doi.org/10.1145/3196709.3196741
[23] Bogdan-Florin Gheran, Radu-Daniel Vatavu, and Jean Vanderdonckt. 2018. Ring x2: Designing Gestures for Smart Rings using Temporal Calculus. In *Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems*. ACM, New York, NY, USA, 117–122. https://doi.org/10.1145/3197391.3205422
[24] Bogdan-Florin Gheran, Santiago Villarreal-Narvaez, Radu-Daniel Vatavu, and Jean Vanderdonckt. 2022. RepliGES and GEStory: Visual Tools for Systematizing and Consolidating Knowledge on User-Defined Gestures. In *Proceedings of the*

ACM International Conference on Advanced Visual Interfaces (AVI '22). ACM, New York, NY, USA, Article 5, 9 pages. https://doi.org/10.1145/3531073.3531112

[25] Omar S. Gómez, Natalia Juristo, and Sira Vegas. 2010. Replications Types in Experimental Disciplines. In *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '10)*. ACM, New York, NY, USA, Article 3, 10 pages. https://doi.org/10.1145/1852786.1852790

[26] Jesús M. González-Barahona and Gregorio Robles. 2012. On the reproducibility of empirical software engineering studies based on data retrieved from development repositories. *Empir. Softw. Eng.* 17, 1-2 (2012), 75–89. https://doi.org/10.1007/S10664-011-9181-9

[27] Steven N. Goodman, Daniele Fanelli, and John P.A. Ioannidis. 2016. What Does Research Reproducibility Mean? *Science Translational Medicine* 8, 341 (June 2016), 341ps12. http://dx.doi.org/10.1126/scitranslmed.aaf5027

[28] Kasper Hornbæk. 2015. We Must Be More Wrong in HCI Research. *Interactions* 22, 6 (2015), 20–21. https://doi.org/10.1145/2833093

[29] Kasper Hornbæk, Søren S. Sander, Javier Andrés Bargas-Avila, and Jakob Grue Simonsen. 2014. Is once enough? on the extent and content of replications in human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3523–3532. https://doi.org/10.1145/2556288.2557004

[30] Ivan Ivanov, Jean Bézivin, and Mehmet Aksit. 2002. Technological Spaces: An Initial Appraisal. In *Proceedings of the 4th International Symposium on Distributed Objects and Applications* (Irvine, USA) *(DOA 2002)*. 1–6.

[31] Peter Kedron, Sarah Bardin, Joseph Holler, Joshua Gilman, Bryant Grady, Megan Seeley, Xin Wang, and Wenxin Yang. 2024. A Framework for Moving Beyond Computational Reproducibility: Lessons from Three Reproductions of Geographical Analyses of COVID-19. *Geographical Analysis* 56, 1 (2024), 163–184. https://doi.org/10.1111/gean.12370

[32] Nathan Magrofuoco, Jorge Luis Pérez-Medina, Paolo Roselli, Jean Vanderdonckt, and Santiago Villarreal. 2019. Eliciting Contact-Based and Contactless Gestures With Radar-Based Sensors. *IEEE Access* 7 (2019), 176982–176997. https://doi.org/10.1109/ACCESS.2019.2951349

[33] Nathan Magrofuoco and Jean Vanderdonckt. 2019. Gelicit: A Cloud Platform for Distributed Gesture Elicitation Studies. *Proc. ACM Hum.-Comput. Interact.* 3, EICS, Article 6 (June 2019), 41 pages. https://doi.org/10.1145/3331148

[34] Rob McCarney, James Warner, Steve Iliffe, Robbert van Haselen, Mark Griffin, and Peter Fisher. 2007. The Hawthorne Effect: a randomised, controlled trial. *BMC Med. Res. Methodol.* 7, 30 (2007). https://doi.org/10.1186/1471-2288-7-30

[35] Manoel G. Mendonça, José C. Maldonado, Maria C.F. de Oliveira, Jeffrey Carver, Sandra C.P.F. Fabbri, Forrest Shull, Guilherme H. Travassos, Erika Nina Höhn, and Victor R. Basili. 2008. A Framework for Software Engineering Experimental Replications. In *Proceedings of the 13th IEEE International Conference on Engineering of Complex Computer Systems*. 203–212. https://doi.org/10.1109/ICECCS.2008.38

[36] Maria Luiza Mondelli, A. Townsend Peterson, and Luiz M. R. Gadelha. 2019. Exploring Reproducibility and FAIR Principles in Data Science Using Ecological Niche Modeling as a Case Study. In *Advances in Conceptual Modeling, Proceedings of ER 2019 Workshops FAIR, MREBA, EmpER, MoBiD, OntoCom, and ER Doctoral Symposium Papers (Lecture Notes in Computer Science, Vol. 11787)*. Springer, Cham, 23–33. https://doi.org/10.1007/978-3-030-34146-6_3

[37] Miguel A. Nacenta, Yemliha Kamber, Yizhou Qiang, and Per Ola Kristensson. 2013. Memorability of pre-designed and user-defined gesture sets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1099–1108. https://doi.org/10.1145/2470654.2466142

[38] Prasad Patil, Roger D. Peng, and Jeffrey T. Leek. 2019. A Visual Tool for Defining Reproducibility and Replicability. *Nat. Hum. Behav.* 3 (2019), 650–652. https://doi.org/10.1038/s41562-019-0629-z

[39] Roger D. Peng. 2011. Reproducible Research in Computational Science. *Science* 334, 6060 (2011), 1226–1227. http://dx.doi.org/10.1126/science.1213847

[40] Darya Pokutnaya, Bruce Childers, Alice E. Arcury-Quandt, Harry Hochheiser, and Willem G. Van Panhuis. 2023. An implementation framework to improve the transparency and reproducibility of computational models of infectious diseases. *PLoS Comput. Biol.* 19, 3 (2023), 1–9. https://doi.org/10.1371/journal.pcbi.1010856

[41] Stefan Pröll, Kristof Meixner, and Andreas Rauber. 2016. Precise Data Identification Services for Long Tail Research Data. http://dx.doi.org/10.6084/M9.FIGSHARE.3847632

[42] Karen Rust, Meethu Malu, Lisa Anthony, and Leah Findlater. 2014. Understanding childdefined gestures and children's mental models for touchscreen tabletop interaction. In *Proceedings of the 2014 Conference on Interaction Design and Children* (Aarhus, Denmark) *(IDC '14)*. ACM, New York, NY, USA, 201–204. https://doi.org/10.1145/2593968.2610452

[43] Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. 2013. Ten Simple Rules for Reproducible Computational Research. *PLoS Computational Biology* 9, 10 (2013), e1003285. https://doi.org/10.1371/journal.pcbi.1003285

[44] Ovidiu-Andrei Schipor and Radu-Daniel Vatavu. 2018. Invisible, Inaudible, and Impalpable: Users' Preferences and Memory Performance for Digital Content in Thin Air. *IEEE Pervasive Computing* 17, 4 (2018), 76–85. https://doi.org/10.1109/MPRV.2018.2873856

[45] Jonathan W. Schooler. 2014. Metascience Could Rescue the "Replication Crisis". *Nature* 515, 9 (2014), 9. http://dx.doi.org/10.1038/515009a

[46] Carmel Shavitt, Anastasia Kuzminykh, Itay Ridel, and Jessica R. Cauchard. 2021. Naturally Together: A Systematic Approach for Multi-User Interaction With Natural Interfaces. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 349, 31 pages. https://doi.org/10.1145/3476090

[47] Arthur Sluÿters, Sébastien Lambot, Jean Vanderdonckt, and Santiago Villarreal-Narvaez. 2024. Analysis of User-Defined Radar-Based Hand Gestures Sensed Through Multiple Materials. *IEEE Access* 12 (2024), 27895–27917. https://doi.org/10.1109/ACCESS.2024.3366667

[48] Nikita Soni, Schuyler Gleaves, Hannah Neff, Sarah Morrison-Smith, Shaghayegh Esmaeili, Ian Mayne, Sayli Bapat, Carrie Schuman, Kathryn A. Stofer, and Lisa Anthony. 2019. Do user-defined gestures for flatscreens generalize to interactive spherical displays for adults and children?. In *Proceedings of the 8th ACM International Symposium on Pervasive Displays* (Palermo, Italy) *(PerDis '19)*. ACM, New York, NY, USA, Article 24, 7 pages. https://doi.org/10.1145/3321335.3324941

[49] Nikita Soni, Schuyler Gleaves, Hannah Neff, Sarah Morrison-Smith, Shaghayegh Esmaeili, Ian Mayne, Sayli Bapat, Carrie Schuman, Kathryn A. Stofer, and Lisa Anthony. 2020. Adults' and Children's Mental Models for Gestural Interactions with Interactive Spherical Displays. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3313831.3376468

[50] Eric W.K. Tsang and Kai-Man Kwan. 1999. Replication and Theory Development in Organizational Science: A Critical Realist Perspective. *The Academy of Management Review* 24, 4 (1999), 759–780. http://dx.doi.org/10.2307/259353

[51] Jean Vanderdonckt, Radu-Daniel Vatavu, Julie Manon, Romain Maddox, Michael Saint-Guillain, Philippe Lefevre, and Jessica J. Márquez. 2025. UX, but on Mars: Exploring User Experience in Extreme Environments with Insights from a Mars Analog Mission. In *Proceedings of the Designing Interactive Systems Conference (DIS '25)*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3715336.3735706

[52] Jean Vanderdonckt, Radu-Daniel Vatavu, and Arthur Sluÿters. 2024. Engineering Touchscreen Input for 3-Way Displays: Taxonomy, Datasets, and Classification. In *Companion Proceedings of the 16th ACM SIGCHI Symposium on Engineering Interactive Computing Systems* (Cagliari, Italy) *(EICS '24 Companion)*. ACM, New York, NY, USA, 57–65. https://doi.org/10.1145/3660515.3661331

[53] Radu-Daniel Vatavu. 2012. User-Defined Gestures for Free-Hand TV Control. In *EuroITV '12*. ACM, NY, USA, 45–48. https://doi.org/10.1145/2325616.2325626

[54] Radu-Daniel Vatavu and Bogdan-Florin Gheran. 2025. Intermanual Deictics: Uncovering Users' Gesture Preferences for Opposite-Arm Referential Input, from Fingers to Shoulder. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, New York, NY, USA, Article 283, 16 pages. https://doi.org/10.1145/3706598.3713474

[55] Radu-Daniel Vatavu and Jacob O. Wobbrock. 2015. Formalizing Agreement Analysis for Elicitation Studies: New Measures, Significance Test, and Toolkit. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. ACM, New York, NY, USA, 1325–1334. https://doi.org/10.1145/2702123.2702223

[56] João Ricardo Nickenig Vissoci, Clarissa Rodrigues Garcia, Luciano de Andrade, Jose Eduardo Santana, Amrapali Zaveri, and Ricardo Pietrobon. 2013. A Framework for Reproducible, Interactive Research: Application to health and social sciences. (2013). https://doi.org/10.48550/arXiv.1304.5688

[57] Max Wilson, Wendy Mackay, Ed Chi, Michael Bernstein, and Jeffrey Nichols. 2012. RepliCHI SIG: From a panel to a new submission venue for replication. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems (CHI EA '12)*. ACM, New York, NY, USA, 1185–1188. https://doi.org/10.1145/2212776.2212419

[58] Jacob O. Wobbrock, Htet Htet Aung, Brandon Rothrock, and Brad A. Myers. 2005. Maximizing the guessability of symbolic input. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (Portland, OR, USA) *(CHI EA '05)*. ACM, New York, NY, USA, 1869–1872. https://doi.org/10.1145/1056808.1057043

[59] Jacob O. Wobbrock, Meredith Ringel Morris, and Andrew D. Wilson. 2009. User-Defined Gestures for Surface Computing. In *CHI '09*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1518701.1518866

[60] Ionuţ-Alexandru Zaiţi, Stefan-Gheorghe Pentiuc, and Radu-Daniel Vatavu. 2015. On Free-Hand TV Control: Experimental Results on User-Elicited Gestures with Leap Motion. *Personal and Ubiquitous Computing* 19, 5-6 (2015), 821–838. https://doi.org/10.1007/s00779-015-0863-y