From Standards to Practice: A Position on Challenges in Operationalizing Software Quality-in-Use

Yvette D. Hastings[1][a] and Ann Marie Reinhold[1,2][b]
[1]*Gianforte School of Computing, Montana State University, Bozeman, Montana, U.S.A.*
[2]*Pacific Northwest National Laboratory, Richland, Washington, U.S.A.*
yvettehastings@montana.edu, reinhold@montana.edu

Keywords: Quality-in-Use, Software Quality Assurance, Empirical Software Validation, Methodological Rigor.

Abstract: Software quality-in-use (QIU) provides a stakeholder centered framework for evaluating whether a software product achieves beneficial outcomes within its specified context of use. Formalized in ISO/IEC 25019, QIU extends beyond traditional usability testing by incorporating broader attributes related to stakeholder beneficialness, acceptability, and freedom from risk. Despite this broader concept, many empirical studies continue to operationalize QIU through usability testing. This narrowing to usability testing limits the comprehensiveness of QIU evaluations. Our position is that shortcomings in recent operationalizations of QIU exhibit weaknesses in construct clarity, methodological rigor, and reporting practices. We present examples of QIU studies that remain usability centric and identify critical methodological gaps that undermine their ability to capture the multifaceted dimensions of QIU. To address these gaps, we propose three recommendations: (1) clarifying QIU constructs and context of use, (2) improving methodological rigor, and (3) strengthening reporting practices. Finally, we outline our ongoing and future work to improve QIU evaluations to better reflect its stakeholder and context centered foundations.

1 INTRODUCTION

Software systems have become a necessary factor in our everyday activities in both personal and professional contexts. As our reliance on software continues to grow, it becomes imperative to evaluate and enhance software quality throughout the software development lifecycle.

Software quality can be evaluated using established frameworks and standards. For example, the ISO/IEC SQuaRE family of standards, particularly the 25010 and 25019 standards, provide structured frameworks used to evaluate the internal, external, and quality-in-use (QIU) aspects of software quality (ISO/IEC 25010:2023(E), 2023; ISO/IEC 25019:2023(E), 2023). While these standards define measurable product characteristics, software success ultimately depends on whether software systems support stakeholders to achieve meaningful goals within their actual contexts of use. Thus, *software quality evaluations must extend beyond product properties to include stakeholder perception, experience, and outcomes*.

The ISO/IEC 25019 standard formalizes a way to capture the stakeholder perspective through the concept of QIU. Here, QIU is defined as "the extent to which the system or product, when used in a specified context of use, satisfies or exceeds stakeholders' needs to achieve specified beneficial goals or outcomes" (ISO/IEC 25019:2023(E), 2023). The ISO/IEC 25019 standard explicitly incorporates multiple stakeholder groups, including direct and indirect users, organizations, and societal elements, and emphasizes context-dependent goal attainment. Under this framework, software quality is not solely about code correctness or interface functionality, but about whether software use produces beneficial, acceptable, and risk-mitigated outcomes.

Despite the availability of standardized QIU frameworks, many studies have operationalized software quality through usability testing. Within software usability engineering (SUE) and human–computer interaction (HCI) research, usability testing provides systematic methods for assessing interaction performance and user experience. Common measures include task success, efficiency, error rates, and satisfaction (Bevan, 2009), often evaluated through controlled or scenario-based studies (Horn-

[a] https://orcid.org/0009-0000-2143-5634
[b] https://orcid.org/0000-0003-0411-3486

bæk, 2006). These methods yield valuable insight into interface-level performance and the effectiveness of interaction design.

However, software usability represents only one component of QIU. In the ISO/IEC 25019 standard, usability is a subcharacteristic of beneficialness. Usability only captures interaction quality rather than the full spectrum of stakeholder outcomes defined under QIU. While effective interaction is a necessary precondition for high QIU, it is not sufficient to demonstrate that stakeholders achieve beneficial, acceptable, and safe outcomes in real-world contexts. Hence, evaluations centered exclusively on usability-centered metrics overlook broader organizational impacts, trust-related concerns, compliance implications, and potential risks associated with software use.

Although QIU has conceptually been articulated since the early works by Bevan (Bevan, 1995) and later formalized in ISO 9241-11, ISO/IEC 25010, and ISO/IEC 25019, empirical studies and industry practice frequently conflate usability testing with QIU in both study design and reporting. As a result, the multidimensional nature of stakeholder-centered software success (i.e., QIU) is often reduced to interface-level performance metrics.

Stated plainly, QIU is more than just usability (Table 1). ***Our position is that the narrow focus on usability limits the ability to capture a holistic understanding of stakeholder perception and decision-relevant outcomes while using software to achieve specific goals***. We expand on this position by drawing on findings from our systematic literature review (Hastings and Reinhold, 2026), reported separately, to identify recurring challenges in how QIU is currently operationalized. Specifically, we analyze how current studies operationalize QIU by focusing on usability testing (i.e., by highlighting recurring weaknesses in design and reporting) while neglecting other aspects of QIU. In short, QIU is not being assessed comprehensively. Here, we propose recommendations to better align empirical practice with QIU's broader goals to measure the stakeholder perceptions around system beneficialness, acceptance, and freedom from risk. Failure to consider all aspects of QIU results in severe consequences, such as high financial costs, legal liability, and loss of life (Section 5; Table 2).

2 BACKGROUND & MOTIVATION

When QIU is operationalized through usability testing metrics alone, evaluations overlook broader QIU dimensions such as stakeholder trust, societal impacts, and software compliance. The conflation between QIU and usability testing originated in SUE and HCI research during the 1970s and 1980s, when the primary aim was to detect user-interface design issues rather than to discover patterns with implications for overall software quality (Nielsen, 1994; Hornbæk, 2006). Consequently, metrics such as task-completion rates and heuristic evaluations became proxies for stakeholder satisfaction. Over time, these proxies fostered an implicit assumption that high usability equates to high QIU.

Quality-in-use emerged as a response to the limited scope of traditional usability assessments focused mainly on user-interface performance. The differentiation between SUE, HCI, and QIU is illustrated in Figure 1, where SUE and HCI are positioned as foundational disciplines focused on system interaction and design, while QIU represents the outcome-oriented framework that encompasses and extends beyond usability testing (ISO/IEC 25019:2023(E), 2023; Bevan, 2009; Fukuzumi, 2022).

When comparing these three disciplines, a key distinction lies in how stakeholder perception is conceptualized and measured. For example, the concept of user experience overlaps between HCI and QIU. However, HCI investigates user experience as a psychological and interactional phenomenon that encompasses user behaviors, cognitive processes, and affective responses, whereas QIU measures experience as a quality outcome. Similarly, SUE integrates usability attributes into the software development lifecycle, whereas QIU evaluates the realized quality of use once the system is deployed in its intended context. These differences are reflected in the standards guiding SUE, HCI, and QIU evaluations practices (i.e., ISO 9241, Nielsen, Mayhew, and ISO/IEC 25019). In short, usability testing evaluates how well users interact with a system, whereas QIU evaluates whether that interaction produces meaningful, trusted, and risk-mitigated outcomes in a real-world context (Table 1).

Even as QIU was codified in international standards, many studies continued to use legacy usability survey instruments (e.g., SUS (System Usability Scale), SUMI (Software Usability Measurement Inventory)) as proxies for QIU constructs (e.g., high user satisfaction equates to high user acceptability and beneficialness) (Bevan, 2009). The convenience

Table 1: Comparison of usability and QIU definitions, focus, key metrics, primary stakeholders, and summary.

Aspect	Usability[1]	Quality-in-use[2]
Definition	The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.	The extent to which the system or product, when used in a specified context of use, satisfies or exceeds stakeholders' needs to achieve specified beneficial goals or outcomes.
Focus	User interaction with the software system (e.g., learnability, ease of use, etc.).	Stakeholder beneficialness, acceptability, and risk mitigation of software in a given context of use.
Key metrics	Effectiveness (task completion rate), efficiency (time on a task), and satisfaction (often user perception), functionality (does the system do what it is supposed to do)	Beneficialness (extent of benefit resulting from the use of a product, system or service; **evaluated as usability** (*i.e., effectiveness, efficiency, & satisfaction*), **accessibility, & suitability**), acceptability (the end user responds favorably to the installation or use of a product; **evaluated as experience, trustworthiness, & compliance**), and freedom from risk (extent to which a product or system **mitigates the potential risk to economic status, human life, health, society, financial values, enterprise activities, or the environment**). **NOTE: Usability is a subattribute of beneficialness** (Figure 1).
Primary stakeholders	End users interacting with a system	Multiple stakeholders, including end users, organizations, or societal elements
Summary	Concerned with how efficiently, effectively, and satisfactorily users can interact with the system to achieve their goals.	Concerned with whether the system as a whole delivers beneficial outcomes, acceptable experiences, and mitigates risk for all stakeholders in its actual context of use.

[1]Definitions shown exactly as presented by Bevan (Bevan, 2009)

[2]Definition shown exactly as presented in the ISO/IEC 25019:2023 standard (ISO/IEC 25019:2023(E), 2023)

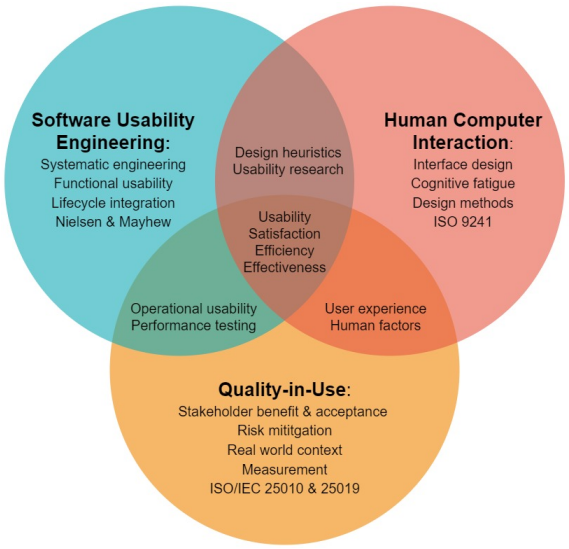

Figure 1: Conceptual differentiation between Software Usability Engineering (blue), Human-Computer Interaction (red), and Quality-in-Use (orange). Software engineering focuses on systematic processes for designing and testing usability attributes. Human-Computer Interaction focuses on interaction design, cognitive factors, and user experience phenomena. Quality-in-Use, as defined in ISO/IEC 25010 and 25019, evaluates stakeholder benefit, acceptance, and risk mitigation efforts arising from software use and outcomes in real contexts of use. Overlapping regions emphasize the shared concerns (e.g., effectiveness, efficiency, satisfaction), while the outer regions highlight distinctions between disciplines. Figure made using Draw.io.

and familiarity of these tools, combined with small-sample, low-cost evaluation norms, has fueled the ongoing conflation (Hertzum, 2016).

Contemporary SUE and HCI evaluation practices also rely on automated analytics or narrowly scoped user studies that emphasize measurable interface metrics (Schaffernak et al., 2025; Verdecchia et al., 2023). While valuable, these approaches overlook the contextual, behavioral, and perceptual dimensions that QIU was designed to capture. This highlights the deeper challenge researchers often face when inheriting frameworks optimized for usability testing rather than assessing human-centered quality outcomes in their context of use.

This persistent misalignment motivates our discussion. To realize the human-centered and evidence-driven intent of QIU, software evaluation must move toward methods that capture how software fulfills stakeholder uses and goals within its specific context of use.

3 CHALLENGES

Although QIU is a comprehensive construct encompassing beneficial outcomes, acceptability, and freedom from risk within a specified context of use in international standards, most QIU evaluations remain narrow in scope and application. Informed by findings from our prior work (Hastings and Reinhold, 2026), where we systematically reviewed 38 stud-

ies evaluating software QIU between the years of 2014-2024, this position paper identifies three recurring challenges: **(1) QIU approached as a usability study**, **(2) methodological shortcomings**, and **(3) missed opportunities from study findings.** We highlight representative evidence from our literature review in the following paragraphs to motivate and support these claims.

Operationalization Challenge 1: QIU Approached as a Usability Study. QIU is defined as a comprehensive construct encompassing both human perception and contextual product outcomes. However, many studies continue to operationalize QIU narrowly through usability-focused evaluations. For example, Khairani et al. (Khairani et al., 2020) frame their contribution in terms of QIU but primarily rely on traditional usability measures (i.e., effectiveness, efficiency, and satisfaction) to assess online education platforms. Similarly, Wulandari et al. (Wulandari et al., 2018) evaluate QIU through interface layout and user feedback on ease of use rather than broader QIU attributes.

This pattern persists across software application domains. Indeed, these studies illustrate a recurring methodological issue. That is, QIU is invoked to justify evaluation frameworks but not operationalized according to its broader definition. Consequently, human-perceived software quality is primarily evaluated through usability constructs, overlooking broader QIU attributes. In practice, attributes such as trustworthiness, accessibility, suitability for use, and freedom from risk are either omitted or treated implicitly, rather than measured and interpreted as QIU outcomes.

Operationalization Challenge 2: Methodological Shortcomings. Although QIU evaluations can be conducted using manual, automated, or hybrid methods, most studies rely on approaches that do not fully support QIU's outcome-oriented intent of demonstrating beneficial outcomes, acceptability, and freedom from risk for relevant stakeholders in a specified context of use.

Manual methods, such as those employed by Wulandari et al. (Wulandari et al., 2018), utilize user testing or heuristic evaluation to identify usability issues, often with a focus on satisfaction or ease of use outcomes (Bevan, 2009). Manual methods provide valuable insights but typically employ usability survey instruments (i.e., SUS or SUMI). These instruments are commonly used to capture aspects related to satisfaction and perceived ease of use, efficiency, or learnability (Bevan, 2009). Additionally, many studies involve small samples and limited replicability (Hertzum, 2016), limiting their ability to generalize QIU beyond interface performance.

Automated approaches, on the other hand, use computational frameworks to evaluate quality attributes without direct stakeholder involvement. For example, Tsuda et al. (Tsuda et al., 2019) automate the measurement of quality characteristics, but emphasize product metrics (e.g., performance, reliability) rather than stakeholder-perceived QIU attributes. While such characteristics can influence QIU, they do not directly show evidence of system beneficialness, acceptability, or freedom from risk for stakeholders in a specified context of use. We argue that while automated evaluations improve reproducibility, they run the risk of oversimplifying the human-centered constructs QIU was designed to represent.

Researchers have attempted to combine the benefits of manual and automated methods using hybrid approaches. For instance, Ben Ayed et al. (Ben Ayed et al., 2016) combine subjective questionnaires with objective log analysis, which provide richer contextual insights relative to single-modality evaluations. However, even hybrid approaches retain a usability-centric focus, using QIU terminology while still emphasizing effectiveness and efficiency over additional attributes. They also focus on usability instruments to determine satisfaction, while using system logs as proxies for behavioral or performance indicators interpreted in relation to QIU attributes. Without explicit construct mapping to QIU, log measures run the risk of being treated as stand-ins for human-centered outcomes. While hybrid methods are employed to characterize QIU and proffer some advantages over manual and automated approaches, as currently deployed, they do not yet fully capture human perception.

Overall, these methodological trends reveal that, while the literature recognizes the importance of QIU, empirical practice remains constrained by usability-driven evaluation traditions. This constraint limits both the generalizability and interpretive depth of QIU as a human-centered construct, which motivates the need for explicit QIU construct mapping, richer context of use specification, and transparent reporting.

Operationalization Challenge 3: Missed Opportunities from Study Findings. A recurring challenge in QIU research is the lack of substantive discussion or follow-up investigation of study findings. Many papers identified user feedback or weaknesses in system design, but did not relate these issues to QIU attributes or acknowledge them as potential limitations to software quality. This limits cross-study synthesis because results cannot be compared or aggregated at the level of QIU concepts.

For example, Delano et al. (Delano et al., 2018) report that participants provided feedback identifying quality weaknesses, specifically noting weaknesses in system implementation and security in their software. However, the authors dismissed these points as "not the focus of this study" and did not address them as future work (Delano et al., 2018). Similarly, Jimenéz-Honrado et al. (Jiménez-Honrado et al., 2024) identified interface-level issues and minor usability concerns but did not connect them to other QIU attributes, such as suitability, trustworthiness, or freedom from risk.

Such limited interpretation of findings suggests that many QIU studies treat user responses as endpoints rather than sources of diagnostic insight. When issues are acknowledged, they are often described qualitatively and detached from the broader QIU framework, which hinders the accumulation of knowledge across studies. This omission undermines one of the key intentions of QIU, which is to evaluate whether stakeholders achieve beneficial goals and outcomes in a specified context of use. A comprehensive treatment of results, including reflection on contextual shortcomings and their implications, is essential to advancing QIU as an evidence-driven evaluation framework.

4 RECOMMENDATIONS

The challenges identified in Section 3 highlight the need for stronger theoretical grounding (such as in the ISO/IEC 25019 standard) and transparency in empirical QIU evaluations. To advance QIU as a rigorous and human-centered construct, we propose three recommendations: **(1) clarify QIU constructs and context of use, (2) improve methodological rigor, and (3) strengthen reporting practices.** These recommendations are designed to align research practices with the intent of QIU, promoting comparability, reproducibility, and interpretive depth across evaluations.

Recommendation 1: Clarify QIU Constructs and Context of Use. We recommend that researchers and practitioners explicitly define how QIU attributes are operationalized within their studies. This includes specifying relevant user groups, task characteristics, and environmental constraints that shape how each QIU construct is interpreted in relation to the context of use, stakeholder type, and intended beneficial outcomes (ISO/IEC 25019:2023(E), 2023). Additionally, studies should include a construct-to-measurement mapping table that links QIU concepts to the corresponding instruments, data sources, and interpretation rules. This process should also involve reconsidering findings initially deemed irrelevant (e.g., trust concerns, compliance issues, or contextual constraints raised by participants), as these may indicate gaps or misalignment in the operationalization of QIU dimensions. By explicitly defining and reporting these mappings, researchers and practitioners reduce conceptual ambiguity, strengthen construct validity, and enable consistent comparison and replication across studies and software domains.

Recommendation 2: Improve Methodological Rigor. To enhance methodological rigor, we recommend that QIU evaluations integrate multiple data sources, instruments, and evaluators. Hybrid designs that combine *subjective* feedback with *objective* log-based measures or performance analytics strengthen methodological rigor and interpretive robustness relative to single-modality evaluations. Such designs strengthen traceability between QIU attributes, measurement instruments, and data interpretation, thereby improving transparency and replicability.

While hybrid designs strengthen rigor through methodological triangulation, manual methods can also advance methodological rigor when systematically and transparently applied. For example, Hastings and Reinhold (Hastings and Reinhold, 2023) operationalized QIU for complex environmental modeling applications. Their goal was to evaluate how effectively such tools support scientists' decision-making during software selection. In this study, they systematically operationalized the ISO/IEC 25010 QIU standard by decomposing its components and subcomponents into measurable metrics and evaluated how each software tool mapped to each QIU criteria. This enhanced traceability between QIU subcharacteristics and the evaluation criteria. This structured manual approach allowed Hastings and Reinhold to recommend tools that supported environmental scientists in achieving their research goals while minimizing financial costs.

Together, these approaches illustrate that methodological rigor in QIU evaluation de-

pends on systematic application of evaluation criteria and transparent procedural design. Careful mapping of QIU concepts to human-centered considerations provides researchers and practitioners with a systematic basis for comparing results across domains and applications. However, even with rigorous methodological design, internal validity remains central, as evaluation outcomes depend on the consistent and transparent application of QIU criteria within the study context.

Recommendation 3: Strengthen Reporting Practices. We recommend that QIU evaluations adopt rigorous reporting standards consistent with empirical software engineering practices. This includes systematically documenting, analyzing, and integrating stakeholder feedback into formal reports to inform necessary improvements in software quality (e.g., using shared reporting templates such as the ISO/IEC 25019 Annexes). Reports should clearly describe study design, data sources, evaluation criteria, and interpretation procedures to facilitate study replication and auditability. Transparent discussion of both positive and negative results improves credibility and enables future researchers to identify gaps, limitations, and opportunities for refinement in existing QIU evaluations. Additionally, reflecting on how contextual or environmental factors that influence stakeholder perceptions–important distinctions in QIU evaluations–strengthens transparency and supports the human-centered objectives of QIU.

In summary, these recommendations emphasize that QIU evaluations benefit from moving beyond usability-centered assessment toward a holistic, transparent, and context-aware empirical approach. By clarifying constructs, improving methodological rigor, and strengthening reporting practices, researchers can better align empirical work with the foundational intent of QIU. Doing so can reduce harmful outcomes from software use.

Collectively, these recommendations strengthen construct validity (through clearer operationalization), internal validity (through rigorous methodological design), and external validity (through transparent reporting and comparability). Without their deliberate adoption and application, empirical QIU research risks continued conceptual dilution and limited cumulative knowledge building.

5 RECOMMENDATIONS IN PRACTICE

To illustrate the utility of heeding the recommendations identified in Section 4, we present a set of real-world scenarios (Table 2). These scenarios demonstrate how the ISO/IEC 25019 standard can be applied to evaluate multiple QIU attributes to support robust, human-centered software quality. Importantly, the scenarios extend beyond traditional usability-focused evaluations by explicitly incorporating the broader QIU attributes defined in the ISO/IEC 25019 standard. The presented scenarios span multiple system domains (e.g., medical, avionics, and public-facing systems) and levels of criticality, including safety-critical systems.

From these scenarios, we illustrate how neglecting specific QIU attributes can lead to severe human, organizational, and societal consequences. Across these domains, failures were not solely usability failures, but breakdowns in beneficialness, acceptability, and freedom from risk. These scenarios are not intended as exhaustive case studies. Rather, they serve as illustrative examples of how QIU can be systematically operationalized to bridge the gap between standards and practice.

To operationalize QIU evaluation in these scenarios, we conceptually map each of the presented system domains to relevant ISO/IEC 25019 characteristics and subcharacteristics and identify how these attributes could be assessed in practice. Assessments include a combination of (1) human-centered evaluation methods (e.g., user testing, surveys, and expert review), (2) system-level validation (e.g., simulation testing, verification of outputs, and failure mode analysis), and (3) context-of-use analysis (e.g., evaluating environmental conditions, user expertise, and operational constraints). This multi-method approach reflects how QIU can be evaluated in practice, where both technical system behavior and user interaction must be considered jointly.

Beyond illustrating how QIU can be applied in practice, the scenarios also highlight the trade-off between implementation costs and realized benefits. In the presented scenarios, we estimated implementation costs based on the level of effort, resources, and system complexity required to conduct QIU evaluations within each context of use. Specifically, costs were categorized as "**Very High**" when evaluations require simulation-based validation, extensive human-in-the-loop testing, or integration with complex sensor or safety-critical systems; "**High**" when evaluations involve substantial code review, expert assessment, and context-of-use testing without full system simulation;

Table 2: Real-world scenarios demonstrating how the ISO/IEC 25019 standard could be operationalized in practice, along with impacted stakeholders and implementation costs and benefits.

Scenario	System Domain	QIU Recommendations	Stakeholders Impacted	Implementation Costs	Implementation Benefits
Therac-25 Radiation Therapy System (Ethics Unwrapped, University of Texas at Austin,)	Medical (safety-critical)	**Beneficialness**: Validate usability and suitability of radiation doses delivered under real clinical conditions; ensure usability through reduced cognitive load. **Acceptability**: Improve experience and trustworthiness through transparent system feedback and error messaging; enforce compliance with medical software standards. **Freedom from Risk**: Explicitly evaluate and mitigate risks to human life through independent safety features and fail-safe mechanisms.	Technicians, physicians, patients, hospitals, software vendor	**High:** Formal verification; safety audits; interface redesign; extended clinical testings	Prevention of fatal overdoses; reduced legal liability; increased trust in medical software; long-term cost savings from avoided recalls and litigation
Boeing V-22 Osprey Flight Control Software (The Air Current,)	Avionics (military; safety-critical)	**Beneficialness**: Assess the usability and suitability of automation in a wide range of operational contexts. **Acceptability**: Improve experience and trustworthiness by aligning automation behavior with pilot mental models; enforce compliance by strengthening traceability between requirements, implementation, and training. **Freedom from Risk**: Systematically evaluate failure modes affecting human life and societal safety.	Pilots and crew, maintenance personnel, military organizations, defense contractors, civilian populations	**Very High:** Simulation-based testing, human-in-the-loop evaluations, multidisciplinary design reviews, training updates	Reduced fatal accidents; improved mission success; increased pilot trust; lower long-term operational and training costs
Boeing 737 MAX MCAS System (Think Reliability,)	Avionics (commercial; safety-critical)	**Beneficialness**: Validate the usability of pilots recovering from automation failures. **Acceptability**: Improve trustworthiness and experience through explainable automation behavior and aligned training; enforce compliance by ensuring certification reflects real operational complexity and context of use. **Freedom from Risk**: Treat single-sensor dependency as unacceptable risk to human life.	Pilots, airline operators, passengers, aviation regulators, manufacturers	**High:** Sensor redundancy, software redesign, pilot retraining, certification rework	Lives saved; avoidance of fleet groundings; restoration of organizational trust; long-term financial stability
Quest Accessibility Kiosks (Kiosk Industry Association,)	Public-facing (non safety-critical)	**Beneficialness**: Improve usability by allowing all users to complete tasks independently; accessibility ensures kiosk interaction supports users with mobility, vision, and reach constraints. **Acceptability**: Improve user experience and compliance with accessibility regulations. **Freedom from Risk**: Mitigate societal and economic risks related to exclusion and legal non-compliance.	Patients, caregivers, deploying organizations, regulators, advocacy groups	**Low–Medium:** Interface redesign, physical hardware adjustments, accessibility testing, staff training	Legal compliance; expanded user base; improved public perception; reduced litigation and loss to reputation risks

and "**Low-Medium**" when evaluations are applied to non safety-critical systems with lower technical complexity but potential organizational or legal implications.

In addition, we determined the implementation benefits by considering the severity and scope of potential consequences associated with QIU failures in each scenario. These benefits range from preventing loss of human life in safety-critical systems to improving user experience, trust, accessibility, and decision-making in public-facing or organizational systems. Thus, benefits are interpreted in terms of risk mitigation, stakeholder protection, and improved system outcomes. These categorizations are intended to provide a comparative, rather than quantitative, assessment of cost and benefit across domains.

While the investment required to evaluate and integrate QIU assessments may be non-trivial, the potential benefits scale significantly with stakeholder impact and system criticality. In safety-critical contexts, neglecting QIU can contribute to cascading failures that place human life at risk, whereas in public-facing or organizational systems the consequences may manifest as exclusion, loss of trust, financial liability, or reputational harm. Thus, the appropriate depth of QIU evaluation should be proportional to system criticality and stakeholder risk exposure. The cost–benefit framing in Table 2 therefore reinforces that QIU should not be treated as an optional usability enhancement, but as a strategic and, in some domains, ethically necessary dimension of software quality.

6 ONGOING & FUTURE WORK

While we can examine systems from historical perspectives and reflect on how they could have been improved, as we did in Section 5, it is equally important to establish a framework that proactively bridges the gap between standards and practice. Our ongoing and future work strives to address this need, as translating QIU standards into actionable evaluations remains a persistent challenge across software domains.

First, we are operationalizing the ISO/IEC 25010 and ISO/IEC 25019 standards across several software application domains, ranging from empirical studies on scientific software applications (Hastings and Reinhold, 2023; Hastings et al., 2026) to planned studies in cybersecurity tools. In our prior work, we developed and applied a QIU-based evaluation framework to support scientific software selection, demonstrating how multiple QIU considerations (e.g., freedom from economic risk, trustworthiness) influence the identification of "best fit" software for achieving stakeholder goals and outcomes (Hastings and Reinhold, 2023). These findings illustrate the importance of looking beyond usability to assess software quality in a holistic and decision-relevant manner.

Second, we are identifying recurring frustrations experienced by end users in their professional software contexts. These challenges are identified through empirical studies of scientific software (Hastings et al., 2026) and will be extended in future work to additional domains, including cybersecurity tools for vulnerability detection and intrusion detection, as well as Software Bill of Materials tools. By reporting on QIU considerations that are not being achieved, particularly those affecting user perception, experience, and outcomes, this work provides targeted recommendations to developers on where software improvements are needed to strengthen QIU.

For future work, we will develop a structured documentation and recommendation framework to support researchers and practitioners in strengthening QIU reporting and dissemination practices. The framework will explicitly record the QIU standard applied, the specific QIU considerations evaluated, the defined context of use, and the observed outcomes of assessment. By requiring this level of construct-level traceability, the approach promotes greater clarity in how QIU attributes are operationalized and interpreted.

Extending beyond documentation, our instantiation of the framework will generate targeted, construct-aligned recommendations to guide software improvement and identify under-evaluated QIU dimensions. Rather than functioning solely as a checklist, it will serve as a scaffold that integrates construct mapping, contextual specification, evaluation-method alignment, and standardized reporting into a coherent and repeatable workflow.

By formalizing these elements, our instantiation of the framework is intended to operationalize Recommendations 1–3 into a repeatable and transparent process. This structure is intended to improve internal validity, enhance reproducibility, and facilitate cross-study comparability. Ultimately, this work advances QIU evaluation from an ad hoc usability-centric practice toward a systematic, synthesis ready methodology that faithfully reflects QIU.

7 CONCLUSION

Quality-in-use represents a critical advancement in how software quality is conceptualized and evaluated. However, limitations exist in how QIU has been operationalized, as legacy usability frameworks that

emphasize interface performance over a contextual human perspective remain the prevailing approach. This ongoing challenge obscures the broader intent of QIU, which is to capture how software supports stakeholders achieving goals and outcomes (ISO/IEC 25019:2023(E), 2023). As illustrated in Table 2, failure to achieve these goals can result in dire consequences, including death.

Through our examination, we identified recurring challenges: (1) the conflation of QIU with usability testing, (2) limited methodological rigor and construct clarity, and (3) restrained discussion of findings and their implications. Collectively, these issues weaken the empirical soundness and interpretive value of current QIU studies, constraining their contribution to the advancement of human-centered software quality evaluation.

To address these challenges, **our position is that future studies prioritize conceptual clarity, methodological rigor, and reporting practices.** Adhering to these recommendations will enhance the credibility and comparability of QIU research. Such practices not only align with the principles of QIU but also reinforce the broader goal of human-centered software evaluation.

To bridge the gap from challenges to practice, we will develop a documentation and recommendation framework, instantiated as a tool. The goal of this tool is to assist researchers and practitioners in moving beyond usability studies labeled as QIU, increase reporting rigor, and generate targeted recommendations for QIU considerations to include when evaluating software.

Ultimately, advancing QIU research requires integrating methodological rigor with software evaluated in its context of use. By grounding evaluations in human perspective while maintaining transparency, researchers and practitioners close the gap between QIU assessments and their original conceptual foundations, producing software quality outcomes that are both measurable and meaningfully reflective of stakeholder perception. Without this shift, QIU risks remaining a terminological extension of usability rather than a distinct, outcome-oriented framework for human-centered software quality.

ACKNOWLEDGMENT

The authors acknowledge the use of ChatGPT to assist in generating ideas for the implementation costs and benefits columns in Table 2.

We thank the Montana State University Software Engineering and Cybersecurity Laboratory for their support of this work. We also thank Dr. C. Izurieta, Dr. D. Reimanis, E. Sheppard, Z. Wadhams, and J. Vang for their review and feedback on this work.

REFERENCES

Ben Ayed, E., Kolski, C., Magdich, R., and Ezzedine, H. (2016). Towards a context based evaluation support system for quality in use assessment of mobile systems. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, page 004350–004355. IEEE Press.

Bevan, N. (1995). Measuring usability as quality of use. *Journal of Software Quality*, 4(2):115–130.

Bevan, N. (2009). Extending quality in use to provide a framework for usability measurement. In *Proceedings of Human-Centered Software Engineering (HCSE)*, pages 13–22, Berlin, Heidelberg. Springer.

Delano, J. D., Jain, H. K., and Sinha, A. P. (2018). System design through the exploration of contemporary web services. In *ACM Transactions on Management Information Systems*, volume 9, pages 1–29, New York, NY, USA. Association for Computing Machinery.

Ethics Unwrapped, University of Texas at Austin. Therac-25 Case Study. Accessed: Dec. 14, 2025.

Fukuzumi, S. (2022). What is the difference between usability in iso 25000 and quality in use? *IEICE Transactions on Information and Systems*, E105-D(10):1705–1713.

Hastings, Y. D., Carver, J. C., Manzi Muneza, A. R., Payn, R. A., Ewing, S. A., Warnat, S., and Reinhold, A. M. (2026). Mining user forums to evaluate quality-in-use of environmental software. In *Proceedings of SoutheastCon 2026*. Accepted for publication; to appear in IEEE Xplore.

Hastings, Y. D. and Reinhold, A. M. (2023). Applying software quality in use standards to improve scientific software selection. *WiPiEC Journal-Works in Progress in Embedded Computing Journal*, 9(2).

Hastings, Y. D. and Reinhold, A. M. (2026). Software quality-in-use: A systematic literature review. In *2026 Intermountain Engineering, Technology and Computing (IETC)*. Accepted for publication; to appear in IEEE Xplore.

Hertzum, M. (2016). Usability testing: too early? too much talking? too many problems? *J. Usability Studies*, 11(3):83–88.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *Int. J. Hum.-Comput. Stud.*, 64(2):79–102.

ISO/IEC 25010:2023(E) (2023). *Systems and software engineering –Systems and software Quality Requirements and Evaluation (SQuaRE) – Product quality model*. International Organization for Standardization, Vernier, Geneva, Switzerland.

ISO/IEC 25019:2023(E) (2023). *Systems and software engineering –Systems and software Quality Requirements and Evaluation (SQuaRE) – Quality-in-use*

model. International Organization for Standardization, Vernier, Geneva, Switzerland.

Jiménez-Honrado, J., Gómez García, J., Costa-Tebar, F., A. Marco, F., Gallud, J. A., and Sebastián Rivera, G. (2024). Progressive web application for storytelling therapy support. In *Proceedings of the XXIV International Conference on Human Computer Interaction*, Interacción '24, New York, NY, USA. Association for Computing Machinery.

Khairani, D., Rosyada, D., Zulkifli, Burhanudin Lubis, A., Daffa Oktriyana, A., and Dewi Herawati Jana, E. (2020). Quality in use measurement of google classroom in online learning. In *2020 8th International Conference on Cyber and IT Service Management (CITSM)*, pages 1–5.

Kiosk Industry Association. Accessibility kiosks: Quest kiosk violates ada. https://kioskindustry.org/accessibility-kiosks-quest-kiosk-violates-ada/. Accessed: 2025-12-12.

Nielsen, J. (1994). *Usability Engineering*. Morgan Kaufmann, Boston, MA.

Schaffernak, H., Moesl, B., Url, P., Victoria Koglbauer, I., and Vorraber, W. (2025). Towards sustainable software quality in use: a review of measures. *Next Research*, 2(3):100680.

The Air Current. V-22 Ospreys Have Serious Unresolved Safety Risks, NAVAIR Review Confirms. Accessed: Dec. 14, 2025.

Think Reliability. Four lessons from the boeing 737 max-8 crashes. https://blog.thinkreliability.com/four-lessons-from-the-boeing-737-max-8-crashes. Accessed: 2025-12-14.

Tsuda, N., Washizaki, H., Honda, K., Nakai, H., Fukazawa, Y., Azuma, M., Komiyama, T., Nakano, T., Suzuki, H., Morita, S., Kojima, K., and Hando, A. (2019). Wsqf: Comprehensive software quality evaluation framework and benchmark based on square. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 312–321. IEEE Press.

Verdecchia, R., Engström, E., Lago, P., Runeson, P., and Song, Q. (2023). Threats to validity in software engineering research: A critical reflection. *Inf. Softw. Technol.*, 164(C).

Wulandari, E., Effendy, V., and Ary Wisudiawan, G. A. (2018). Modeling user interface of first-aid application game using user centered design (ucd) method. In *2018 6th International Conference on Information and Communication Technology (ICoICT)*, pages 354–359. IEEE Press.