Towards Usable E-Health

A Systematic Review of Usability Questionnaires

Vanessa E. C. Sousa1; Karen Dunn Lopez2

1Department of Health Systems Science, University of Illinois at Chicago, Chicago, IL, USA

Keywords
Usability, e-health, questionnaires, systematic review

Summary

Background: The use of e-health can lead to several positive outcomes. However, the potential for e-health to improve healthcare is partially dependent on its ease of use. In order to determine the usability for any technology, rigorously developed and appropriate measures must be chosen.

Objectives: To identify psychometrically tested questionnaires that measure usability of e-health tools, and to appraise their generalizability, attributes coverage, and quality.

Methods: We conducted a systematic review of studies that measured usability of e-health tools using four databases (Scopus, PubMed, CINAHL, and HAPI). Non-primary research, studies that did not report measures, studies with children or people with cognitive limitations, and studies about assistive devices or medical equipment were systematically excluded. Two authors independently extracted information including: questionnaire name, number of questions, scoring method, item generation, and psychometrics using a data extraction tool with pre-established categories and a quality appraisal scoring table.

Results: Using a broad search strategy, 5,558 potentially relevant papers were identified. After removing duplicates and applying exclusion criteria, 35 articles remained that used 15 unique questionnaires. From the 15 questionnaires, only 5 were general enough to be used across studies. Usability attributes covered by the questionnaires were: learnability (15), efficiency (12), and satisfaction (11). Memorability (1) was the least covered attribute. Quality appraisal showed that face/content (14) and construct (7) validity were the most frequent types of validity assessed. All questionnaires reported reliability measurement. Some questionnaires scored low in the quality appraisal for the following reasons: limited validity testing (7), small sample size (3), no reporting of user centeredness (9) or feasibility estimates of time, effort, and expense (7).

Conclusions: Existing questionnaires provide a foundation for research on e-health usability. However, future research is needed to broaden the coverage of the usability attributes and psychometric properties of the available questionnaires.

Correspondence to:
Vanessa E. C. Sousa, PhD, MSN
University of Illinois at Chicago, College of Nursing
Department of Health Systems Science
845 South Damen St.
Chicago, IL 60612
Email: vanessaemille@gmail.com
Phone: 773–814–0517

Appl Clin Inform 2017; 8: 470–490
https://doi.org/10.4338/ACI-2016-10-R-0170
received: October 10, 2016
accepted: February 26, 2017
published: May 10, 2017

Citation: Sousa VEC, Lopez KD. Towards usable e-health: A systematic review of usability questionnaires. Appl Clin Inform 2017; 8: 470–490
https://doi.org/10.4338/ACI-2016-10-R-0170

Funding
This study was supported in part by funding from The National Council for Scientific and Technological Development (CNPq), Brasilia, Brazil.
1. Background and Significance

E-Health is a broad term that refers to a variety of technologies that facilitate healthcare, such as electronic communication among patients, providers and other stakeholders, electronic health systems and electronically distributed health services, wireless and mobile technologies for health care, telemedicine and telehealth and electronic health information exchange [1]. There has been exponential growth in the interest, funding, development and use of e-health in recent years [2–5]. Their use has led to a wide range of positive outcomes including improved: diabetes control outcomes [6], asthma lung functions [7], medication adherence [8], smoking cessation [9], sexually transmitted infection testing [10], tuberculosis cure rate [11], and reduced HIV viral load [12]. However, not all e-health demonstrate positive outcomes [13]. It is likely that even for e-health based in strong evidence based content, if the technology is difficult to use, the overall effectiveness on patient outcomes will be thwarted. In order to determine the ease of use (usability) for any new technology, rigorously developed and appropriate measures must be chosen [14, 15].

The term “usability” refers to set of concepts. Although usability is a frequently used term, it is inconsistently defined by both the research community [16] and standards organizations. The International Standards Organization (ISO) number 9241 defines usability as “the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use” [17]. Unfortunately, ISO has developed multiple usability standards each with differing terms and labels for similar characteristics [16].

In the absence of a clear consensus, we chose to use Jakob Nielsen’s five usability attributes: learnability, efficiency, memorability, error rate and recovery, and satisfaction (Figure 1) [18]. Dr. Nielsen is highly regarded in the field of human computer interaction [19] and his usability attributes are the foundation for voluminous number of usability studies [20–25] including those of e-health [26–29]. Both ISO 9241 [17] and Nielsen’s [18] definition share the concepts of efficiency and satisfaction, but the key advantage of Nielsen’s definition over ISO’s 9241 is the clarity and specificity of the additional three concepts of learnability, memorability and error rate and recovery over the more general concept of efficiency in ISO’s definition.

[32–35] Usability together with utility (defined as whether a system provides the features a user needs) [18] comprise overall usefulness of a technology. Usability is so critical to the effectiveness of e-health that even applications with high utility may become unlikely to be accepted if usability is poor [30–33]. Beyond the problem of poor acceptance, e-health with compromised usability can also be harmful for patients. For example, medication errors that are facilitated by electronic health records (EHRs) with compromised efficiency affects the clinicians’ ability to find needed test results, view coherent listings of medications, or review active problems which can result in delayed or incorrect treatment [34]. A computerized provider order entry system with too many windows can increase the likelihood of selecting wrong medications [35]. Other studies have shown that healthcare information technology with fragmented data (e.g. the need to open many windows to access patient data) leads to the loss of situation awareness, which compromises the quality of care and the ability to recognize and prevent impending events [36–38].

Given the importance of usability a wide range of methods for both developing usable technologies as well as assessing the usability of developed technologies have been created. Methods range from user inspection methods such as heuristic evaluations [39], qualitative think aloud interviews [39, 40], formal evaluations frameworks [41], simulated testing [42] and questionnaires. Although questionnaires have less depth than data yielded from qualitative analysis and may not be suitable when technologies are in the early stages of development, questionnaires play an important role in usability assessment. A well-tested questionnaire is generally much less expensive than using qualitative methods. In addition, unlike qualitative data, many questionnaires can be analyzed using predictive statistical analysis that can advance our understanding of technology use, acceptance and consequences.

Despite its importance, quantitative usability assessment of e-health applications have been hampered by two major problems. The first problem is that the concept of usability is often misapplied and not clearly understood. For example, some studies have reported methods of usability measurement that do not include any of its attributes [43] or include only partial assessments that do not
capture the whole meaning of usability [44–50]. It is also common to find usability being used as a synonym of acceptability [51, 52] or utility [53–56], further confounding the assessment of usability.

The second major problem is that many studies use usability measures that are very specific to an individual technology, or use questionnaires that lack psychometrics such as reliability (consistency of the measurement process) or validity (measurement of what it is supposed to be measured) [57–60]. Although we recognize that a usability assessment needs to consider specific or unique components of some e-health, we also believe that generalizable measures of usability that can be used across e-health types can be useful to the advancement of usability science. For example, to examine underuse of a patient portals, an analysis would be more robust if they measured both the patient and clinician portal interface using the same measure. There are also several benefits to having a generalizable usability measures to improve EHR usability. This includes comparative benchmarks for EHR usability across organizations, within organizations following upgrades, availability of comparable usability data prior to EHR purchase [61] and creating incentives for vendors to compete on usability performance [61]. In addition, having a generalizable measure of usability could be included as a way to operationalize Section 4002 (Transparent Reporting on Usability, Security, And Functionality) of the recently signed into law 21st Century Cures Act [62].

In sum, appropriate usability measurement is essential to give any technology implementation the best chance of success and to identify potential mediators of e-health engagement and health-related outcomes. However, questionnaires that have been used over the years to assess usability of e-health have not been systematically described or examined to guide the choice of strong measures for research in this area. The purpose of this review is to identify psychometrically tested questionnaires to measure usability of e-health tools, and to appraise their generalizability, attributes coverage, and quality.

2. Methods

2.1. Search Strategy and Study Selection

We conducted a review of four databases: Scopus, MEDLINE via PubMed, the Cumulative Index to Nursing Allied Health (CINAHL) via Ebscohost, and Health and Psychosocial Instruments (HaPI), from October 6th to November 3rd, 2015. Search terms included: “usability”, “survey”, “measure”, “instrument”, and “questionnaire”, combined by the Boolean operators AND and OR. We used different search strategies in each database to facilitate the retrieval of studies that included measures of usability developed with users of e-health tools (Supplementary online File 1). Our search yielded 5,558 articles. After 2,206 duplicates were removed, a sample of 3,352 articles were considered for inclusion.

Papers were identified, screened and selected based on specific inclusion and exclusion criteria applied in three stages: 1) title review (2282 studies excluded), 2) abstract review (877 studies excluded), and 3) full article review (158 studies excluded) (Figure 2). Title exclusions rules included articles that were: not in English, not primary research, not inclusive of usability measures, not e-health tool related, conducted with children and people with cognitive limitations, and testing of assistive devices and medical equipment. We chose to exclude studies with children and people with cognitive limitations as subjects because questionnaires used with these populations may not be applicable to general adult users of e-health. We excluded assistive devices and medical equipment because of their specific features (e.g. user satisfaction with feedback for moving away from obstacles), which cannot be compared with questions applied to other types of technology.

Abstract based exclusion rules included articles that: used fully qualitative approach or met previous title exclusion rules. Full article based exclusion rules included articles that: did not assess any usability attribute, with usability measures unavailable, or with psychometrics unavailable in the article or citations. This provided a final sample of 35 articles.

Each round of exclusion (title, abstract and full article) was individually assessed by two authors using a sample of 5% of the articles. When disagreements arose, they were resolved by discussion. The use of two assessors continued for 3–6 rounds until inter-rater agreement was established for
each step at 85% (90% for titles review after 3 rounds, 85% for abstracts review after 6 rounds, and 97% for full articles review after 1 round).

2.2. Data Extraction and Analysis

Two investigators extracted data from papers using a data extraction tool built using Google Forms with variables, categories and definitions. We derived categories for variables using our general knowledge of usability, testing the categories using a subsample of articles, and adding new labels inductively, as necessary to accommodate important information that did not fall into any of the existing categories. Data entered was automatically stored in an online spreadsheet and assessed for agreement reaching the goal of >85% at the first round.

We extracted data into two stages. Stage 1 extracted general data from each of the 35 studies that met our inclusion and exclusion rules. This included: authors, place and year of publication, type of e-health technology evaluated, questionnaire used and origin of the questions. Because some of the questionnaires were used in more than one article, Stage 2 focused on the 15 unique questionnaires identified in the sample of 35 articles. For this stage, we extracted more specific data about the questionnaires’ development and psychometric testing. When necessary, we extracted information about the questionnaires’ development and psychometric assessment from the reference list of the original studies, or by the questionnaires’ original authors.

2.3. Generalizability, Attributes Coverage and Quality Appraisal

Generalizability was assessed by one question asking if the questionnaires’ items were generic or technology-related, i.e., whether the questionnaires included items referring to specific features that are not common across e-health applications. Attributes coverage was evaluated for each usability attribute (learnability, efficiency, memorability, error rate and recovery, and satisfaction) using Nielsen’s definitions [18] (Figure 1). The quality assessment was comprised of the five criteria used in Hesselink, Kuis, Pijnenburg and Wollersheim [63]: validity, reliability, user centeredness, sample size, and feasibility (Supplementary online File 2). We did not include the criteria of responsiveness (the ability of a questionnaire to detect important changes over time) from the original tool because none of the studies assessed usability over time. The total score possible for each questionnaire is 10.

Definitions from the literature [57] for the specific types of validity and reliability can be seen in Supplementary online File 3. From these definitions, we emphasize that user centeredness refers to “the inclusion of users’ opinions or views to define or modify items.” Thus, we carefully searched for information in each study about the participation of potential users during the questionnaires’ development stage to rate this construct and based our ratings on whether this information was present or not in the published article.

3. Results

Our inclusion and exclusion rules yielded 35 unique articles that used 15 unique questionnaires.

3.1. Descriptive Analysis of Studies

The majority of the 35 studies were conducted in the United States (n=13) [64–76] and in European countries (n=8) [77–84]. The primary subjects were health workers (n=17) [59, 65, 66, 70–72, 75, 78, 80, 81, 85–91] and patients (n=11) [69, 73, 74, 76, 78, 79, 84, 91–94]. The types of e-health tools tested in the studies were comprised of: clinician-focused systems (n=15) [59, 65, 66, 70–72, 75, 80, 81, 85–88, 90, 95], patient-focused systems (n=10) [64, 67, 69, 74, 76, 83, 91–94], wellness/fitness applications (n=4) [77, 79, 84, 96], electronic surveys (n=3) [73, 78, 97], and digital learning objects (n=3) [68, 82, 89] (Supplementary online File 4).
3.2 Descriptive Analysis of Questionnaires

We identified 15 unique questionnaires across the 35 studies that measured individual perceptions of usability. The number of items in the questionnaires ranged from 3 to 38. Fourteen out of 15 questionnaires were Likert type, and 1 used a visual analog scale [92]. Eleven of the questionnaires [59, 82, 85, 86, 88, 90, 96, 98–101] had subscales. Most questionnaires were derived from empirical studies (pilot testing with human subjects) [59, 82, 85, 90, 98–100, 102, 103]. The others were derived from theories or models and from the literature (Table 1).

3.3. Generalizability

Generalizability assessments revealed that 10 questionnaires were created by the studies’ authors specifically for their e-health and contain several items that may be too specific to be generalized (e.g. “The built-in hot keys on the CPOE system facilitate the prescription of physician orders”). Only 5 questionnaires include items that could potentially be applied across different types of e-health tools: the System Usability Scale (SUS) [103], the Questionnaire for User Interaction Satisfaction (QUIS) [100], the After-Scenario Questionnaire (ASQ) [104], the Post-Study System Usability Questionnaire (PSSUQ) [99], and the Computer System Usability Questionnaire (CSUQ) [105].

3.4. Attributes Coverage

Each questionnaire was evaluated in terms of usability attributes coverage. Learnability was the most covered usability attribute (all questionnaires). The least assessed usability attributes were: (1) error rate/recovery, that was included in only 6 questionnaires: PSSUQ [99], CSUQ [105], QUIS [100], Lee, Mills, Bausell and Lu [88], Oztekin, Kong and Uysal [82], and Peikari, Shah, Zakaria, Yasir and Elhissi [90]; and (2) memorability, that was accessed only by Oztekin, Kong and Uysal [82]. The 4 questionnaires that had the highest attribute coverage were: QUIS [100], Lee, Mills, Bausell and Lu [88], PSSUQ [99], and CSUQ [105] (Figure 3 and Figure 4). All but 2 questionnaires [59, 78, 82, 85, 86, 88, 90, 92, 96, 98–101] also included items that measured the e-healths’ utility.

3.5. Quality Assessment of Studies

An overview of the quality appraisal is shown in Table 2. Quality scores ranged from 1–7 of a possible 10 points, and the average score was 4.1 (SD 1.9). The maximum score (7 points) was achieved by only 2 questionnaires: SUS [103] and Peikari, Shah, Zakaria, Yasir and Elhissi [90].

Face or content validity, often used as interchangeable terms, were addressed in 14 questionnaires. Construct validity, performed by a series of hypothesis tests to determine if the measure reflects the unobservable constructs [106], was established by exploratory factor analysis in 7 questionnaires [82, 86, 98–100, 102, 107], and by confirmatory factor analysis in 2 questionnaires [82, 90]. Criterion validity, assessed by correlating the new measure with a well-established or “gold standard” measure [106], was addressed by only 1 study [102].

Reliability, a measure of reproducibility [106], was assessed for all questionnaires by a Cronbach’s α coefficient or by composite reliability. All questionnaires had an acceptable or high reliability based on the 0.70 threshold [108]. Inter-item correlations were reported for only 1 questionnaire [109]. The questionnaire’s correlation coefficients were weak for some items, but strong for others, based on a 0.50 threshold [110]. Test-retest reliability, a determination of the consistency of the responses over time, was assessed for 1 questionnaire [89] and resulted in a correlation coefficient above the minimum threshold of 0.70 [111].

In addition to classic psychometric evaluation of the measure’s quality, Hesselink and colleagues [63] quality assessment method also includes sample size, feasibility and user centeredness. Three studies [82, 86, 99] had small samples (below 5 participants per item) and 5 studies [90, 98, 100, 102, 109] had acceptable samples (5–10 participants per item) based on Kass and Tinsley [112] guidelines. The remaining studies did not perform factor analysis so there was not a sample size standard to evaluate. Feasibility, related to respondent burden, was assessed in 8 studies, but this was measured in disparate ways. Six studies reported difficulties perceived by users [59, 85, 88, 90, 100,
4. Discussion

The aim of this study was to appraise the generalizability, attributes coverage, and quality of questionnaires used to assess usability of e-health tools. We were surprised to find that none of the questionnaires cover all of the usability attributes or achieved the highest possible quality score using Hesselink, Kuis, Pijnenburg and Wollersheim [63]. However, by combining the generalizability, attributes coverage, and quality criteria, we believe the strongest of the currently available the questionnaires are the SUS, the QUIS, the PSSUQ, and the CSUQ. Although the SUS does not cover efficiency, memorability or errors, it is a widely-used questionnaire [113] with general questions that can be applied to a wide range of e-health. In addition, the SUS achieved the highest quality score of the identified questionnaires. The QUIS, the PSSUQ, and the CSUQ also include measures that are general to many types of e-health and have the advantage of covering additional usability attributes (efficiency and errors) when compared with SUS. However, we emphasize that researchers should define which usability measures are the best fit for the intent of the study, technology being assessed, and context of use. For example, an EHR developer may be more concerned about creating a system with low error rate than user satisfaction, while the developer of an “optional” technology (e.g. patient portal or exercise tracking) is likely to need a measure of satisfaction. In addition, we acknowledge that there may be specific factors in some e-health (e.g. size and weight of a hand-held consumer focused EKG device) that need to be measured along with the general usability attributes. In these instances, we recommend using a high-quality general usability questionnaire along with well-tested and e-health specific questions.

There were notable weaknesses found across many of the questionnaires. We were surprised to learn that most the questionnaires (10 out of 15) were specific to a single technology, as their items focus on aspects specific to unique e-health tools. Others have noted the challenges associated with a decision to use specific or generalizable usability measures and advocate for modifying items to address specific systems and user tasks under evaluation [114, 115]. Although this approach can be helpful, it can affect the research subjects’ comprehension of the questions and can change the psychometric properties of the questionnaire. Thus, questionnaires having adapted or modified items need to be tested before use [116].

We believe there is value to having the community of usability scientists use common questionnaires that allow comparisons across technology types. For examples knowing the usability ratings of commerce focused applications versus e-health can help set benchmarks for raising the standards of e-health usability. In addition, having common questionnaires can help pin point usability issues in an underused e-health. One potential questionnaire to promote for such purposes is the SUS. The SUS is the only usability questionnaire we identified that allows researchers to “grade” their e-health on the familiar A-F grade range often used in education. It has also has been cited in more than 1,200 publications and translated in eight languages becoming one of the most widely used usability questionnaires [103, 107, 113, 117]. This is not to say that SUS does not have weaknesses. In particular, we believe this measurement tool could be strengthened by further validity testing and by expanding the usability coverage to include efficiency, memorability and errors.

We were a bit disheartened to find that no single questionnaire enables the assessment of all usability attributes defined by Nielsen [18]. Most questionnaires have items covering only one or two aspects of usability (learnability and efficiency were the most common attributes), while other important aspects (such as memorability and error rate/recovery) are left behind. Incomplete or inconsistent assessments of the usability of e-health technologies are problematic and can be harmful. For example, a technology can be subjectively pleasing but have poor learnability and memorability, requiring extra mental effort from its users [118]. An EHR can be easy to learn but at the same time can require the distribution of information over several screens (poor efficiency), resulting in increased workload and documentation time [119]. A computerized physician order entry system that make it difficult to detect errors can increase the probability of prescribing errors [35]. These
examples also serve to illustrate importance of usability to e-health and the fact that adoption is not a measure of success of e-health, especially for technologies that are mandated by organizations or strongly encourage through national policies.

We found that most questionnaires measured utility along with usability, and were especially surprised to find that these terms were used interchangeably. Our search also yielded studies (excluded using our exclusion rules) that did not measure any of the usability attribute despite using this term either in their titles, abstracts, or within the text [120, 121]. Together these findings suggests that despite several decades of measurement, usability is an immature concept that is not consistently defined [122] or universally understood.

The quality appraisal showed that most questionnaires lack robust validity testing despite being widely used. While all questionnaires have been accessed for internal consistency and face/content validity (assessed subjectively and considered the weakest form of validity), other types of validity (content, construct and criterion) are missing. Face and content validity are not enough to ensure that a questionnaire is valid because they do not necessarily refer to what the test actually measures, but rather to a cursory judgment of what the questions appear to measure [123]. Lacking of objective measures of validity can result in invalid or questionable study findings, since the variables may not accurately measure the underlying theoretical construct [124, 125].

Cronbach’s alpha estimates of reliability were acceptable or high for most studies, thus indicating that these questionnaires have good internal consistency (assessment of how well individual items correlate between each other in measuring the same construct) [57]. Cronbach’s alpha tends to be the most frequently used estimate of reliability mainly because the other estimates require experimental or quasi-experimental designs, which are difficult, time-consuming and expensive [126, 127]. Rather than an insurmountable weakness of the quality of the existing questionnaires, our findings endorse the idea that there is still opportunity to further develop the reliability and validity of these measures. In particular, we think that Heikkinen, Suomi, Jaaskelainen, Kaljonen, Leino-Kilpi and Salantera [92] using a visual analog scale, merits further development. One advantage of this type of measure is the ability to analyze the scores as ease of continuous variables that opens opportunities to conduct more advanced analytic methods when compared to ordered Likert categories that were used in most the questionnaires.

The quality appraisal also revealed a somewhat surprisingly finding: we identified a lack of user centeredness in all but 6 studies [59, 85, 88, 90, 100, 103]. Incorporating users’ views to define or modify questionnaire’s items is important in any kind of research that involves human perceptions [128]. If the goals of the questionnaire and each of its items are not clear, the answer’s given by the users will not reflect what they really think, yielding invalid results.

Feasibility estimates (time, effort, and expenses involved in producing/using the questionnaire) were also lacking in most studies and were insufficiently described in others. Feasibility is an important concern because even highly reliable questionnaires can be too long, causing unwillingness to complete, mistakes or invalid answers [128]. At the same time that respondent burden is a problem, short questionnaires that are too brief (like the ASQ) can lead to insufficient coverage of the attributes intended to be measured [129].

Finally, we acknowledge that both quantitative and qualitative methods play important roles in technology development and improvement. While quantitative methods have the advantages of being generally inexpensive and more suitable for large sample studies, qualitative methods (like think-aloud protocols) are useful to provide details about specific sources of problems that quantitative measures cannot usually match. In addition, qualitative assessments provide information about user behaviors, routines, and a variety of other information that is essential to deliver a product that actually fit into a user’s needs or desires [130]. Ideally, both qualitative and quantitative approaches should be applied in the design or improvement of technologies.

### 4.1. Limitations

Despite our best efforts for rigor in this systematic review, we note a few limitations. First, given the focus of the review on quantitative usability questionnaires, we did not include all methods of usability evaluation. This means that several well important approaches such as heuristic review methods [39], [41], and think aloud [39] were beyond the scope of this review. Second, although we used
specific terms to retrieve measures of usability, we found that several of the questionnaires also in-
cluded questions about utility in their measures. Although, the concept of utility is related to usabil-
ity, we do not consider this review to systematically address utility measures. Because of the import-
ance of utility in combination with usability to user's technology acceptance, we think a review of
utility measures in e-health is an important direction for future studies. Third, although we com-
bined different terms to increase sensitivity we may have missed some articles with potentially rel-
evant questionnaires because of the search strategy we used. For example, our exclusion criteria may
have led to exclusion of research published in other languages than English. Finally, the scoring
method selected for quality appraisal of the studies may have underestimated the quality of some
studies by favoring certain methodological characteristics (e.g. having more than one type of valid-
ity/reliability) over others. However, we were unable to identify a suitable alternative method to ap-
praise the quality of questionnaires.

5. Conclusions

Poor usability of e-health affect the chances of achieving both adoption and positive outcomes. This
systematic review provides a synthesis of the quality of questionnaires that are currently available for
usability measurement of e-health. We found that usability is often misunderstood, misapplied, and
partially assessed, and that many researchers have used usability and utility as interchangeable
terms. Although there are weaknesses in the existing questionnaires, efforts to include the strongest
and most effective measures of usability in research could be the key to delivering the promise of
e-health.

Clinical Relevance Statement

This article provides evidence on the generalizability, attributes coverage, and quality of question-
naires that have been used in research to measure usability of several types of e-health technology.
This synthesis can help researchers in choosing the best measures of usability based on their intent
and technology purposes. The study also contributes to a better understanding of concepts that are
essential for developing and implementing usable and safe technologies in healthcare.

Questions

1. An electronic health record is used to retrieve a patient's laboratory results over time. In this hypo-
thetical system, each laboratory result window only displays the results for one day, so the user must
open several windows to access information for multiple dates to make clinical decisions. Which us-
ability attribute is compromised?
   A) Efficiency
   B) Learnability
   C) Memorability
   D) Few errors

The correct answer is A: Efficiency. Efficiency is concerned with how quickly users can perform
tasks once they learned how to use a system. Thus, a system that requires accessing several windows
to perform a task is an inefficient system. The major problem that results from inefficient systems is
that users' productivity may be reduced. You can rule out learnability because it is related to how
easy it is for users to accomplish tasks, using a system for the first time. Memorability can also be
ruled since it is related to the ability of a user to reestablish proficiency after a period not using a sys-
tem. Errors can be ruled since it is concerned with how many errors users make, error severity or
ease of error recovery [16].
2. What is the risk inherent in using a usability questionnaire that has not undergone formal validity testing?
A) Low internal consistency
B) Lack of user-centeredness
C) Inaccurate measurement of usability
D) Inability to compare technologies usability quantitatively

The correct answer is: C inaccurate measurement of usability. Validity testing is a systematic process hypothesis testing methods that concerned with whether a questionnaire measures what it intends to measure [93]. Without testing to determine whether the new questionnaire measures usability, it is possible that the questionnaire items targets a different concept. We found that several of the questionnaires included in this review appear to measure utility not usability. You can rule out internal consistency, often measured by the Cronbach alpha coefficient, because internal consistency refers to whether the items on a scale are homogenous [93]. You can also rule out user centeredness because validity testing does not concern itself with user centeredness. You can rule out option D (Inability to compare technologies usability quantitatively), because it is possible that many researchers may use the same questionnaire that has not been systematically validated allowing them to compare across technologies on the possibly invalid measure.

Conflict of Interest
The authors declare that they have no conflicts of interest in the research.

Human Subjects Protections
Human subjects were not included in the project.
**Fig. 1** Nielsen’s attributes and definitions [18]

- **Utility**: Whether a system provides the features the user needs
- **Usability**: How easy and pleasant user interfaces are to use
- **Learnability**: Ease of learning the functionality and the behavior of the system
- **Efficiency**: Level of attainable productivity of the user after he has learned the system
- **Memorability**: Ease of remembering the system functionality, so that the casual user can return to the system after a period of non-use, without needing to learn again how to use it
- **Few errors**: Capability of the system to support users in making less error during the use of the system, and in case they make errors, to let them easily recover
- **Satisfaction**: Measure of how pleasant the design is to use

**Fig. 2** PRISMA flow diagram for article inclusion

- Records identified in 4 databases ($n = 5558$)
  - Titles screened after 2206 duplicates removed ($n = 3352$)
  - Abstracts assessed for eligibility ($n = 1070$)
  - Full-text articles assessed for eligibility ($n = 193$)
- Additional records identified through other sources ($n = 0$)
  - Records excluded ($n = 2282$)
    - Not English ($n = 205$)
    - Not primary research ($n = 315$)
    - Not inclusive of measures ($n = 118$)
    - Not e-health tool related ($n = 462$)
    - Children or people with cognitive limitations ($n = 540$)
    - Testing of assistive devices ($n = 335$)
    - Testing of medical equipment ($n = 307$)
  - Records excluded ($n = 877$)
    - Fully qualitative approach ($n = 287$)
    - Met previous exclusion rules ($n = 590$)
  - Full-text articles excluded ($n = 158$)
    - Not assessed usability attributes ($n = 10$)
    - Measures unavailable ($n = 67$)
    - Psychometrics unavailable in the article/citations ($n = 81$)
- Studies included ($n = 35$)
  - Unique questionnaires identified ($n = 15$)
### Fig. 3 Attributes covered by each questionnaire

<table>
<thead>
<tr>
<th>Questionnaire name</th>
<th>Usability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Learnability</td>
</tr>
<tr>
<td>After-Scenario Questionnaire (ASQ)*</td>
<td></td>
</tr>
<tr>
<td>Computer System Usability Questionnaire (CSUQ)*</td>
<td></td>
</tr>
<tr>
<td>Post-Study System Usability Questionnaire (PSSUQ)*</td>
<td></td>
</tr>
<tr>
<td>Questionnaire for User Interaction Satisfaction (QUIS)*</td>
<td></td>
</tr>
<tr>
<td>System Usability Scale (SUS)*</td>
<td></td>
</tr>
<tr>
<td>Albu, Atack, and Srivastava (2015), Not named</td>
<td></td>
</tr>
<tr>
<td>Fritz et al. (2012), Not named</td>
<td></td>
</tr>
<tr>
<td>Hao et al. (2013), Not named</td>
<td></td>
</tr>
<tr>
<td>Heikkinen et al. (2010), Not named</td>
<td></td>
</tr>
<tr>
<td>Huang and Lee (2011), Not named</td>
<td></td>
</tr>
<tr>
<td>Lee et al. (2008), Not named</td>
<td></td>
</tr>
<tr>
<td>Oztekin, Kong, and Uysal (2010), Not named</td>
<td></td>
</tr>
<tr>
<td>Peikari et al. (2015), Not named</td>
<td></td>
</tr>
<tr>
<td>Wilkinson et al. (2004), Not named</td>
<td></td>
</tr>
<tr>
<td>Yui et al. (2012), Not named</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>15</td>
</tr>
</tbody>
</table>

*All items generalizable across e-health
### Table 1  Characteristics of the questionnaires.

<table>
<thead>
<tr>
<th>Questionnaire name</th>
<th>Number of questions and scoring</th>
<th>Subscales</th>
<th>Item generation</th>
</tr>
</thead>
<tbody>
<tr>
<td>After-Scenario Questionnaire (ASQ) [104]</td>
<td>3; 7-point Likert Scale ('strongly agree' to 'strongly disagree') and N/A</td>
<td>-</td>
<td>Empirical study</td>
</tr>
<tr>
<td>Computer System Usability Questionnaire (CSUQ) [105]</td>
<td>19; 7-point Likert Scale ('strongly agree' to 'strongly disagree') and N/A</td>
<td>System usefulness (8); Information quality (7); Interface quality (3)*</td>
<td>Empirical study</td>
</tr>
<tr>
<td>Post-Study System Usability Questionnaire (PSSUQ) [99]</td>
<td>19; 7-point Likert Scale ('strongly agree' to 'strongly disagree') and N/A</td>
<td>System usefulness (7); Information quality (6); Interface quality (3)f</td>
<td>Empirical study</td>
</tr>
<tr>
<td>Questionnaire for User Interaction Satisfaction (QUIS) [100]</td>
<td>27; 10-point Likert Scale (several adjectives positioned from negative to positive) and N/A</td>
<td>Overall reaction to the software (6); Screen (4); Terminology and system information (6); Learning (6); System capabilities (5)</td>
<td>Empirical study</td>
</tr>
<tr>
<td>System Usability Scale (SUS) [103]</td>
<td>10; 5-point Likert Scale ('strongly disagree' to 'strongly agree')</td>
<td>-</td>
<td>Empirical study</td>
</tr>
<tr>
<td>Albu, Atack, and Srivastava (2015), Not named [96]</td>
<td>12; 5-point Likert Scale ('strongly disagree' to 'strongly agree')</td>
<td>Ease of use (8); Usefulness (4)</td>
<td>Theory/Model</td>
</tr>
<tr>
<td>Fritz and colleagues (2012), Not named [78]</td>
<td>17; 5-point Likert Scale ('strongly disagree' to 'strongly agree')</td>
<td>-</td>
<td>Literature</td>
</tr>
<tr>
<td>Hao and colleagues (2013), Not named [85]</td>
<td>23; 5-point Likert Scale ('very satisfied' to 'very dissatisfied')</td>
<td>System Operation (5); System Function (4); Decision Support (5); System Efficiency (5); Overall Performance (4)</td>
<td>Empirical study</td>
</tr>
<tr>
<td>Heikkinen and colleagues (2010), Not named [92]</td>
<td>12; Visual analogic scale (100 mm)</td>
<td>-</td>
<td>Literature</td>
</tr>
<tr>
<td>Huang and Lee (2011), Not named [86]</td>
<td>30; 4-point Likert Scale ('no idea or disagreement' to 'absolute understanding or agreement')</td>
<td>Program design (8); Function (7); Efficiency (5); General satisfaction (10)</td>
<td>Literature</td>
</tr>
<tr>
<td>Lee and colleagues (2008), Not named [88]</td>
<td>30; 4-point Likert Scale ('strongly disagree' to 'strongly agree')</td>
<td>Patient care (6); Nursing efficiency (6); Education/training (6); Usability (6); Usage benefit (6)</td>
<td>Literature</td>
</tr>
<tr>
<td>Questionnaire name</td>
<td>Number of questions and scoring</td>
<td>Subscales</td>
<td>Item generation</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>--------------------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>Oztekin, Kong, and Uysal (2010), Not named [82]</td>
<td>36; 5-point Likert Scale ('strongly disagree' to 'strongly agree')</td>
<td>Error prevention (3); Visibility (3); Flexibility (2); Course management (4); Interactivity, feedback and help (3); Accessibility (3); Consistency and functionality (3); Assessment strategy (3); Memorability (4); Completeness (3); Aesthetics (2); Reducing redundancy (3)</td>
<td>Literature Empirical study</td>
</tr>
<tr>
<td>Peikari and colleagues (2015), Not named [90]</td>
<td>17; 5-point Likert Scale ('strongly disagree' to 'strongly agree')</td>
<td>Consistency (4); Ease of use (3); Error prevention (3); Information quality (3); Formative items (4)</td>
<td>Literature Empirical study</td>
</tr>
<tr>
<td>Wilkinson and colleagues (2004), Not named [101]</td>
<td>38; 5-point Likert Scale ('strongly disagree' to 'strongly agree')</td>
<td>Computer use (7); Computer learning (5); Distance learning (4); Overall course evaluation (7); Fulfilment of learning outcomes (1); Course support (7); Utility of the course material (7)</td>
<td>NR‡</td>
</tr>
<tr>
<td>Yui and colleagues (2012), Not named [59]</td>
<td>28; 4-point Likert Scale ('strongly disagree' to 'strongly agree')</td>
<td>Interface design (6); Operation functions (11); Effectiveness (5); Satisfaction (6)</td>
<td>Literature Empirical study</td>
</tr>
</tbody>
</table>

*Factor analysis of CSUQ showed that Item 19 loaded in two factor, thus this item was not included in any subscale. †The original version of PSSUQ did not contain Item 8. Factor analysis of PSSUQ showed that Items 15 and 19 loaded in two factor, thus they are not part of any subscale. ‡NR: Not Reported
Table 2  Quality appraisal of the questionnaires.

<table>
<thead>
<tr>
<th>Questionnaire name</th>
<th>Validity</th>
<th>Reliability</th>
<th>User centeredness</th>
<th>Sample size*</th>
<th>Feasibility</th>
<th>Quality score</th>
</tr>
</thead>
<tbody>
<tr>
<td>After-Scenario Questionnaire (ASQ) [104]</td>
<td>Face/Content validity Established by experts Construct validity PCA: 8 factors accounted for 94% of total variance Criterion validity Correlation between ASQ and scenario success: -.40 (p&lt;.01)</td>
<td>Cronbach’s α α &gt; 0.9</td>
<td>NR†</td>
<td>Sufficient</td>
<td>Time Participants took less time to complete ASQ than PSSUQ (amount of time not reported)</td>
<td>06</td>
</tr>
<tr>
<td>Computer System Usability Questionnaire (CSUQ) [105]</td>
<td>Face/Content validity Established by experts Construct validity PCA: 3 factors accounted for 98.6% of the total variance</td>
<td>Cronbach’s α α = 0.95</td>
<td>NR</td>
<td>Sufficient</td>
<td>NR†</td>
<td>05</td>
</tr>
<tr>
<td>Post-Study System Usability Questionnaire (PSSUQ) [99]</td>
<td>Face/Content validity Established by experts Construct validity PCA: 3 factors accounted for 87% of the total variance</td>
<td>Cronbach’s α α = 0.97</td>
<td>NR</td>
<td>Low</td>
<td>Time Participants needed about 10 min to complete PSSUQ</td>
<td>05</td>
</tr>
<tr>
<td>Questionnaire for User Interaction Satisfaction (QUIS) [100]</td>
<td>Face/Content validity Established by experts Construct validity PCA: 4 latent factors resulted from the factor analysis</td>
<td>Cronbach’s α α = 0.94</td>
<td>Included user’s feedback</td>
<td>Sufficient</td>
<td>Perceived difficulty</td>
<td>05</td>
</tr>
<tr>
<td>System Usability Scale (SUS) [109]</td>
<td>Face/Content validity Established by people with different occupations Construct validity PCA: 2 factors accounted for 56–58% of the total variance</td>
<td>Cronbach’s α α = 0.91§ Inter-item correlation 0.34–0.69 Included user’s feedback</td>
<td>Sufficient</td>
<td>Perceived difficulty</td>
<td>07</td>
<td></td>
</tr>
<tr>
<td>Albu, Atack, and Srivastava (2015), Not named [96]</td>
<td>NR</td>
<td>Cronbach’s α α = 0.86</td>
<td>NR</td>
<td>-</td>
<td>NR</td>
<td>01</td>
</tr>
<tr>
<td>Fritz and colleagues (2012), Not named [78]</td>
<td>Face/Content validity Established by experts</td>
<td>Cronbach’s α α = 0.84</td>
<td>NR</td>
<td>-</td>
<td>NR</td>
<td>03</td>
</tr>
<tr>
<td>Hao and colleagues (2013), Not named [85]</td>
<td>Face/Content validity Established by experts</td>
<td>Cronbach’s α α = 0.91</td>
<td>Included user’s feedback</td>
<td>-</td>
<td>Perceived difficulty Training needs</td>
<td>04</td>
</tr>
<tr>
<td>Heikkinen and colleagues (2010), Not named [92]</td>
<td>Face/Content validity Established by experts</td>
<td>Cronbach’s α α = 0.84 to 0.87</td>
<td>NR</td>
<td>-</td>
<td>NR</td>
<td>03</td>
</tr>
<tr>
<td>Huang and Lee (2011), Not named [86]</td>
<td>Face/Content validity Established by experts Construct validity PCA: 3 factors accounted for 98.6% of the total variance</td>
<td>Cronbach’s α α = 0.80</td>
<td>NR</td>
<td>Low</td>
<td>NR</td>
<td>03</td>
</tr>
<tr>
<td>Questionnaire name</td>
<td>Validity</td>
<td>Reliability</td>
<td>User centredness</td>
<td>Sample size*</td>
<td>Feasibility</td>
<td>Quality score</td>
</tr>
<tr>
<td>-------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
<td>-----------------------------------------</td>
<td>------------------</td>
<td>--------------</td>
<td>-------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Lee and colleagues (2008), Not named [88]</td>
<td>Face/Content validity Established by experts</td>
<td>Cronbach’s $\alpha$ $\alpha = 0.83$ to $0.87$</td>
<td>Included user’s feedback</td>
<td>-</td>
<td>Perceived difficulty</td>
<td>06</td>
</tr>
<tr>
<td>Oztekin, Kong, and Uysal (2010), Not named [82]</td>
<td>Face/Content validity Established by the study authors Construct validity PCA: 12 factors accounted for 65.63% of the total variance CFA: Composite Reliability: 0.7; Average Variance: 0.5</td>
<td>Composite reliability CR = 0.83 (Error prevention); 0.81 (Visibility); 0.73 (Flexibility); 0.89 (Management); 0.74 (Interactivity); 0.79 (Accessibility); 0.69 (Consistency)</td>
<td>NR</td>
<td>Low</td>
<td>NR</td>
<td>03</td>
</tr>
<tr>
<td>Peikari and colleagues (2015), Not named [90]</td>
<td>Face/Content validity Established by experts Construct validity CFA: Factor loadings: 0.80 to 0.87 ($p&lt;0.001$, $t=15.38$); Average Variance: &gt;0.5</td>
<td>Cronbach’s $\alpha$ $\alpha = 0.79$ (Information quality); 0.82 (Ease of use); 0.86 (Consistency); 0.78 (Error prevention)</td>
<td>Included user’s feedback</td>
<td>Sufficient</td>
<td>Perceived difficulty</td>
<td>07</td>
</tr>
<tr>
<td>Wilkinson and colleagues (2004), Not named [101]</td>
<td>Face/Content validity Established by experts</td>
<td>Cronbach’s $\alpha$ $\alpha = 0.84$ (Computer use); 0.69 (Computer learning); 0.69 (Distance learning); 0.76 (Course evaluation); 0.91 to 0.94 (Learning outcomes); 0.87 (Course support); 0.75 (Utility) Test–retest reliability $r=0.81$</td>
<td>NR</td>
<td>-</td>
<td>NR</td>
<td>03</td>
</tr>
<tr>
<td>Yui and colleagues (2012), Not named [30]</td>
<td>Face/Content validity Established by experts</td>
<td>Cronbach’s $\alpha$ $\alpha = 0.93$</td>
<td>Included user’s feedback</td>
<td>-</td>
<td>Perceived difficulty</td>
<td>06</td>
</tr>
</tbody>
</table>

*The sample size quality criterion was assessed only for studies who included factor analysis. †PCA: Principal Component Analysis; CFA: Confirmatory Factor Analysis. ‡NR: Not reported. §For the System Usability Scale, Cronbach’s $\bar{\alpha}$ value was extracted from Bangor, Kortum, and Miller (2008) using 2,324 cases, and PCA was extracted from Lewis and Sauro (2009) using Bangor, Kortum, and Miller (2008) data + 324 cases.
References


22. Yan J, El Ahmad, A. S. Usability of CAPTCHA or usability issues in CAPTCHA design. 4th symposium on Usable privacy and security; 2008; Pittsburgh, PA: ACM.


27. Lu JM, Hsu YL, Lu CH, Hsu PE, Chen YW, Wang JA. Development of a telepresence robot to rebuild the physical face-to-face interaction in remote interpersonal communication. International Design Alliance (IDA) Congress Education Congress; October, 24–26th; Taipei, Taiwan; 2011.


62. 21st Century Cures Act, H.R.6. To accelerate the discovery, development, and delivery of 21st century cures; and for other purposes; 2015.
65. Im EO, Chee W. Editors. Evaluation of the decision support computer program for cancer pain management. Oncology nursing forum 2006; Pittsburgh, PA: Oncology Nursing Society.


