CHAPTER 38

USABILITY AND USER EXPERIENCE: DESIGN AND EVALUATION

James R. Lewis

MeasuringU

Delray Beach, Florida

Jeff Sauro

MeasuringU

Denver, Colorado

1 INTRODUCTION

Usability and User Experience (UX) are important concepts in the design and evaluation of products or systems intended for human use (Lewis, 2014; Sauro & Lewis, 2016; Vredenburg, Isensee, & Righi, 2002; Vredenburg, Mao, Smith, & Carey, 2002). Historically, the goal of usability engineering has been to develop products that are objectively effective, efficient, and with which users will be satisfied (ISO, 1998). More recently, usability engineering efforts have expanded their scope beyond the classic definition of usability to UX, which, depending on the specific context of use, includes attention to emotional factors such as pleasure, beauty, and trust (Diefenbach, Kolb, & Hassenzahl, 2014; Hassenzahl, Platz, Burmester, & Lehner, 2000; Jordan, 2002; Oliveira, Alhinho, Rita, & Dhillon, 2017; Safar & Turner, 2005; Tractinsky, Katz, & Ikar, 2000).

The aims of this chapter are the following:

- Briefly introduce the fundamentals of design for usability and UX, focusing on the application of science, art, and craft to their principled design.
- Review the major methods of usability assessment, focusing on usability testing.
- Discuss the various standardized questionnaires that are currently available for the assessment of different aspects of UX.

1.1 What Is Usability?

The term *usability* came into general use in the early 1980s. Related terms from that time were *user friendliness* and *ease of use*, which *usability* has since displaced in professional and technical writing on the topic (Bevan, Kirakowski, & Maissel, 1991). Well before the 1980s, a refrigerator advertisement in the *Palm Beach Post* from March 8, 1936 cited usability as a key feature (S. Isensee, personal communication, January 17, 2010). The earliest scientific publication (of which we are aware) to include the word *usability* in its title was Bennett (1979) "The Commercial Impact of Usability in Interactive Systems."

It is the nature of language that words come into use with fluid definitions. Ten years after the first scientific use of the term *usability*, Shackel (1990, p. 31) wrote, "one of the most important issues is that there is, as yet, no generally agreed definition of usability and its measurement." Eight years later, Gray and Salzman (1998, p. 242) stated: "Attempts to derive a clear and crisp definition of usability can be aptly compared to attempts to nail a blob of Jell-O to the wall." Twenty years after Shackel, according to Alonso-Ríos et al. (2010, p. 53) "A major obstacle to the implantation of User-Centered Design in the real world is the fact that no precise definition of the concept of usability exists that is widely accepted and applied in practice."

There are several reasons why it has been so difficult to define usability. Usability is not a property of a person or thing. There is no thermometer-like instrument that can provide an absolute measurement of the usability of a product (Dumas, 2003; Hertzum, 2010; Hornbæk, 2006). Usability is an emergent property that depends on the interactions among users, products, tasks, and environments.

Introducing a theme that will reappear in several parts of this chapter, there are two major conceptions of usability. These dual conceptions have contributed to the difficulty of achieving a single agreed-upon definition. One conception is that the primary focus of usability engineering should be on measurements related to the accomplishment of global task goals (summative, or measurement-based, evaluation, both objective and subjective). The other conception is that practitioners should focus on the detection and elimination of usability problems (formative, or diagnostic, evaluation).

The first (summative) conception has led to a variety of similar definitions of usability, some embodied in current standards (which, to date, have emphasized summative evaluation). For example (Bevan et al., 1991, p. 652):

> The current MUSiC definition of usability is: the ease of use and acceptability of a system or product for a particular class of users carrying out specific tasks in a specific environment; where "ease of use" affects user performance and satisfaction, and "acceptability" affects whether or not the product is used.

Usability is 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" (American National Standards Institute (ANSI), 2001, p. 3; International Organization for Standardization (ISO), 1998, p. 2). As defined in ISO 9126-1, usability is one of several software characteristics that contribute to quality in use (in addition to functionality, reliability, efficiency, maintainability, and portability), and Bevan (2009) has recommended including flexibility and safety along with traditional summative conceptions of usability in a more complete quality-of-use model. The quality in use integrated measurement (QUIM) scheme of Seffah et al. (2006) includes 10 factors, 26 subfactors, and 127

specific metrics. Winter et al. (2008) proposed a two-dimensional model of usability that associates a large number of system properties with user activities. Alonso-Rios et al. (2010) published a preliminary taxonomy for the concept of usability that includes traditional and nontraditional elements, organized under the primary factors of Knowability, Operability, Efficiency, Robustness, Safety, and Subjective Satisfaction.

These attempts to provide a more comprehensive definition of usability have yet to undergo statistical testing to confirm their defined structures. An initial meta-analysis of correlations among prototypical summative prototypical usability metrics (effectiveness, efficiency, and satisfaction) that used published scientific studies from the human–computer interaction (HCI) literature found generally weak correlations among the different metrics (Hornbæk & Law, 2007). A replication using data from a large set of industrial usability studies, however, found strong correlations among prototypical usability metrics measured at the task level, with principal-components and factor analyses that provided statistical evidence for the underlying construct of usability with clear underlying objective (effectiveness, efficiency) and subjective (task-level satisfaction, test-level satisfaction) factors (Sauro & Lewis, 2009). One of the earliest formative definitions of usability (ease of use) is from Chapanis (1981, p. 3):

> Although it is not easy to measure "ease of use. it is easy to measure difficulties that people have in using something. Difficulties and errors can be identified, classified, counted, and measured. So my premise is that ease of use is inversely proportional to the number and severity of difficulties people have in using software. There are, of course, other measures that have been used to assess ease of use, but I think the weight of the evidence will support the conclusion that these other dependent measures are correlated with the number and severity of difficulties.

Practitioners in industrial settings generally use both conceptualizations of usability during iterative design. Any iterative method must include a stopping rule to prevent infinite iterations. In the real world, resource constraints and deadlines can dictate the stopping rule (although this rule is valid only if there is a reasonable expectation that undiscovered problems will not lead to drastic consequences). In an ideal setting, the first conception of usability can act as a stopping rule for the second. Setting aside, for now, the question of where quantitative goals come from, the goals associated with the first conception of usability can define when to stop the iterative process of the discovery and resolution of usability problems. This combination is not a new concept. In one of the earliest published descriptions of iterative design, Al-Awar et al. (1981, p. 31) wrote:

> Our methodology is strictly empirical. You write a program, test it on the target population, find out what's wrong with it, and revise it. The cycle of test–rewrite is repeated over and over until a satisfactory level of performance is reached. Revisions are based on the performance, that is, the difficulties typical users have in going through the program.

1.2 What Is User Experience (UX)?

The concept of UX casts a broad net over all of the experiential aspects of use, primarily subjective experience (Bargas-Avila & Hornbæk, 2011; Hassenzahl & Tractinsky, 2006; Hertzum & Clemmensen, 2012; Jordan, 2002; McCarthy & Wright, 2004). Due to the growing interest in the broad concept of UX among industrial practitioners, the Usability Professionals Association (UPA), formed in 1991, changed its name to the User Experience Professionals Association (UXPA) in 2012. It is important not to confuse this with discussions of the effect of the amount of experience people have with products or systems, where common metrics are years of use or amount of daily/weekly use (we will refer to this as "product experience" in this chapter). The primary focus of UX measurement is on the emotional consequences of use and resulting behaviors (Lewis, Brown, & Mayes, 2015; Lewis & Mayes, 2014; Saariluoma & Jokinen, 2014).

Before 2000, the primary focus of industrial practitioners working on the development of products/systems for human use was on classical usability, assessing the extent to which designs led to, for example, successful and rapid task completion and high levels of satisfaction or perceived usability. In 2000 and 2001, Marc Hassenzahl and colleagues began to publish research on a distinction between classical usability, which they called pragmatic usability, and hedonic usability, defined by a set of semantic differential items such as interesting-boring and impressive-nondescript (Hassenzahl, 2001, 2018; Hassenzahl et al., 2000). Since then, he has continued to conduct influential research in this area, exploring other drivers of perceived usability and developing the AttrakDiff questionnaire for the assessment of a variety of aspects of the user experience (Hassenzahl et al., 2015). Within twelve years of these initial publications, Diefenbach et al. (2014) were able to find 151 publications that discussed hedonic usability as an aspect of interactive products. Hertzum and Clemmensen (2012) found in interviews with 24 usability practitioners from three different countries (China, Denmark, and India) a distinction between utilitarian and experiential constructs in their work, with experiential constructs covering a broader range than those defined in Hassenzahl's original definition of hedonic usability.

It is interesting that similar changes occurred in the business field of management of information systems (MIS) over a similar time period with regard to the Technology Acceptance Model (TAM). As part of his doctoral dissertation at MIT, Fred Davis developed the first incarnation of the TAM over three decades ago, roughly the same time as the first appearance of the System Usability Scale (SUS, a widely used measure of perceived usability), publishing the model in 1989 (Davis, 1989). Having a reliable and valid measure that could explain and predict usage would be valuable for both software vendors and information technology managers. A number of studies support the validity of the TAM and its satisfactory explanation of end-user system usage (Wu, Chen, & Lin, 2007).

In the first version of TAM, Davis (1989) reported that the key drivers of initial usage were perceived ease-of-use and perceived usefulness, which affected, in order, attitudes toward use, behavioral intention to use, then actual usage. In an attempt to improve the prediction of use, Venkatesh (2000) published the TAM 2 which added more constructs to the model, including some management-related social psychological constructs (self-efficacy and intrinsic motivation) and emotional constructs (computer anxiety, perceived enjoyment). This was followed by TAM 3 (Venkatesh & Bala, 2008), which modeled the determinants of perceived ease-of-use with constructs such as computer self-efficacy, perception of external control, computer anxiety, computer playfulness, perceived enjoyment, and objective usability. These extensions to the original TAM model show the increasing desire to explain the adoption (or lack thereof) of technology and to define and measure the many external variables affecting the original drivers of perceived usefulness and perceived ease-of-use. There is a clear connection between UX and TAM in the relationship between perceived ease-of-use and perceived usability (Lewis, 2018a), both in predicting future use and assessing current usage experience.

Using a modified version of TAM 1 in which participants independently rated their experience with three products (two US groups, one Slovenian), Lah et al. (2020) found that substituting the SUS (perceived usability) for the TAM 1 perceived ease-of-use factor in regression models predicting overall experience and likelihood-to-recommend had no significant effect on the standardized regression coefficients. In other words, from a statistical perspective, SUS and the TAM construct of perceived ease-of-use appeared to be measuring the same thing.

Michalco et al. (2015) conducted research on the relationship between expectations and user experience, finding that prior expectation significantly affected ratings of user experience, with the effect depending on whether expectation was positively or negatively disconfirmed. In another study informed by the relationship between UX and TAM, Aranyi and van Schaik (2015) explored their application to the domain of online news sites, basing their work on the components of user experience model of Thüring and Mahlke (2007). In that model, “user experience is gained in the course of interacting with a technical

device" (Thüring & Mahlke, 2007, p. 262), with system properties, user characteristics, and task/context affecting, though interaction, perception of instrumental qualities (e.g., perceived usability, perceived usefulness), perception of non-instrumental qualities (e.g., aesthetics, novelty), and emotional reactions (e.g., subjective feelings, motor expressions, physiological reactions), leading to final appraisals of the system (e.g., overall judgment, likelihood to recommend, likelihood to switch to a competitor). Aranyi and van Schaik found this model worked well and concluded "We look forward to seeing the CUE [components of user experience] model being applied as a framework in future UX-modeling research in information science and technology, and other domains" (p. 2485).

In summary, research related to the user experience beyond classical, instrumental usability has a history of about two decades. As with the core concept of usability, there is variability in the details of how different researchers have conceptualized UX. Despite this variation in details, there is considerable agreement that UX is an extension of classical usability in which instrumental attributes such as efficiency and effectiveness are still important, but primarily to the extent that they influence emotional outcomes such as satisfaction, trust, and perceived beauty, with resulting effects on outcome behaviors such as repeat purchases and recommendation to others.

2 FUNDAMENTALS OF DESIGN FOR USABILITY AND UX

There are many resources that provide design guidelines for specific types of users, products, and contexts. This includes other chapters in this handbook, style guides published for operating systems, and books dedicated to the design of, for example, voice user interfaces (Lewis, 2011b), mobile experiences (Ballard, 2007), web design (Krug, 2014), and online forms (Jarrett & Gaffney, 2009). In this chapter we take a step back from detailed design and review some higher-level aspects of design for usability and UX.

The first principle of design for usability and UX is to base principles on science, art, and craft. Scientific and engineering disciplines that can inform UX design include:

- Psychology (e.g., cognitive, social, psycholinguistics)
- Human factors engineering (HFE)
- Human–computer interaction (HCI)
- Linguistics (especially pragmatics)
- Communication theory

- Market research/Service science (e.g., e-Service, self-service technologies)

The translation of scientific findings to design is a major part of HFE and HCI (Gillan & Schvaneveldt, 1999; Lewis, 2011a). For example, Fitts' law, which models the time required to acquire a target based on the distance to the target and its size, and the Hick-Hyman law, which models the time required to make decisions based on the number of alternatives, can be combined to inform the design of soft keyboard layouts to maximize input speed when typing with a stylus (Soukoreff & MacKenzie, 1995). Scientific studies of human vision, audition, and touch can inform the design of visual, auditory, and tactile displays (Proctor & Proctor, 2006).

It would be nice to have solid scientific evidence to guide every aspect of design, but that is not a realistic expectation. Where there is no science to inform design, it is necessary to rely on art and craft (instantiated, for example, as design principles or heuristics) for the remaining design decisions. Artistic practices include user-centered design and design thinking, visual and auditory design, and writing (especially technical and script writing). Craft includes using prior experience to guide design and the codification of leading practice in style guides.

2.1 Iterative Design

The basic idea of iterative design is that it is possible to improve initial designs by iterating design and evaluation processes, with each design step informed by the previous evaluation step. Early accounts of iteration applied to product design came from Alphonse Chapanis and his students (Al-Awar et al., 1981; Chapanis, 1981; Kelley, 1984) and had an almost immediate influence on product development practices at IBM (Kennedy, 1982; Lewis, 1982) and other companies, notably Xerox (Smith et al., 1982) and Apple (Williams, 1983). Shortly thereafter, John Gould and his associates at the IBM T. J. Watson Research Center began publishing influential papers on usability testing and iterative design (Gould & Boies, 1983; Gould & Lewis, 1984; Gould et al., 1987; Gould, 1988), as did Whiteside et al. (1988) at DEC (Baecker, 2008; Dumas, 2007).

The driving force that separated iterative design/evaluation from the standard protocols of experimental psychology was the need to modify early product designs as rapidly as possible (as opposed to the scientific goal of developing and testing competing theoretical hypotheses). As Al-Awar et al. (1981, p. 33) reported: "Although this procedure [iterative usability test, redesign, and retest] may seem unsystematic and unstructured, our experience has been that there is a surprising amount of consistency in what subjects report. Difficulties are not random or whimsical. They do form patterns."

When difficulties of use become apparent during the early stages of iterative design, it is hard to justify continuing to ask test participants to perform the test tasks. There are ethical concerns with intentionally frustrating participants who are using a product with known flaws that the design team can and will correct. There are economic concerns with the time wasted by watching participants who are encountering and recovering from known error-producing situations. Furthermore, any delay in updating the product delays the potential discovery of problems associated with the update or problems whose discovery was blocked by the presence of the known flaws. For these reasons, the earlier you are in the design cycle, the more rapidly you should iterate the cycles of test and design.

2.2 User-Centered Design/Design Thinking

Before design iteration can be used to improve a design, there must be a design to iterate. User-centered design (UCD) and design thinking are methods used to produce initial designs, after which they typically use iteration for design improvement (Curedale, 2019; Følstad, Law, & Hornbaek, 2012; Still & Crane, 2017, Vredenburg, Mao, et al., 2002). UCD emerged in the 1980s to extend usability engineering by including an early focus on involving users in pre-design activities to understand user needs, then continuing to include users in evaluations of early design prototypes. Design thinking emerged from professional design practice in the late 1960s, becoming popular in commercial design in the 1990s (Barbaroux, 2016). Although they have different roots, the practices are similar in their focus in user involvement in design and the use of iteration (Karat & Karat, 2003).

The first appearance of "UCD" was in Norman and Draper's (1986) book, *User Centered System Design*. They included "System" in part because they wanted the initials to match the university at which Don Norman taught at the time, the University of California at San Diego (UCSD). Over time, the "S" got dropped because UCD was applicable to design in general, not just system design (Gulliksen et al., 2003). The generic UCD process includes the following steps (Vredenburg, 2003):

- *Market definition:* Determine who will use the offering.
- *Task analysis*: Determine what people will need or want to do with it.
- *Competitive evaluation*: Find out who or what the offering is competing with and develop benchmarks for the assessment of design success.
- *Initial design and walkthrough*: Prototype an initial design and walk through the expected tasks.

- *Iterative design evaluation and validation*: Take the final version of the prototype from the previous step and begin iterative design and usability testing with target users.
- *Benchmark assessment*: Use the previously defined benchmarks to determine whether the design goals have been achieved or if there is a need to continue iterating.

Bruce Archer is generally credited with being the first to publish the term "design thinking" in 1965, with adaptation to business-oriented designs by the company IDEO in 1991 (Barbaroux, 2016). Similar to the current state of UCD, different writers have proposed different design thinking processes (Allanwood & Beare, 2019). A common five-step design thinking process is:

- *Empathy*: Develop empathy with potential users through interviews and ethnographic activities.
- *Synthesis*: Working from users' viewpoints, define the problem you need to solve.
- *Ideation*: Generate a large number of ideas that might solve the problem and encourage innovative thinking.
- *Prototyping*: Evaluate the most promising ideas with prototypes, keeping them as simple as possible while still sufficient to assess how well people from the target audience can use it to complete target tasks.
- *Iterative test and redesign*: Use iteration to refine the idea, monitoring user success and allowing them to critique the design during each iteration.

One of the ways in which specific UCD and design thinking processes differ is in the roles users play. As Karat and Karat (2003, p. 539) wrote:

> There is general agreement that this goal [usability] is achieved through the involvement of potential users in system design. In this we feel we must be somewhat less specific about what role users play in the process than some argue for. For example, in the participatory design community, approaches have been developed to enable the users to take active roles in many design activities. In the context in which these techniques were developed (Scandinavian countries with strong labor unions), users have the right to design their work environments. It is likely that techniques derived from this experience might need to be modified to fit use contexts that are different. System design is ultimately a partnership between developer and user, and the level of partnership between user and developer is a factor that will vary.

2.3 Service Design

Service design is a relatively new area of design for usability and UX practitioners. A major application of information technology is providing service to clients, where clients might be customers who pay for services directly or citizens who pay for services through fees or taxation (Larson, 2008). Service science (Lusch, Vargo, & O'Brien, 2007; Lusch, Vargo, & Wessels, 2008; Pitkänen Virtanen, & Kemppinen, 2008; Spohrer & Maglio, 2008) is an interdisciplinary area of study focused on systematic innovation in service. A key concept of service science is that payment for performance defines service, as opposed to payment for physical goods. Other attributes of service are that it is time-perishable, created and used at the same time, and includes a client participating in the coproduction of value.

As work in a service system evolves, there is a tendency to shift focus from human talent to technology, culminating in self-service. In self-service systems, the balance of investment of time in deriving value from the service has largely shifted to the client who is seeking service (Rowley, 2006). With this shift in responsibility, it is important to design systems that enhance customer efficiency, which then leads to enhanced customer attraction and retention (Xue & Harker, 2002). Four high-level design principles, based on the attributes of effective human service agents (Balentine & Morgan, 2001) are:

- Assume the client is busy.
- Be efficient when communication is good.
- Be helpful when progress is slower.
- Be courteous and polite, rarely apologize, and never blame the client.

The early 2000s saw a significant increase in market research of the drivers and inhibitors of customer satisfaction with self-service technologies. In a landmark paper, Meuter et al. (2000) reported a critical incident study of more than 800 self-service incidents gathered with an Internet survey. The technologies included IVR, Internet, and kiosks. Tasks included customer service (such as telephone banking, flight information, and order status), transactions (such as telephone banking and prescription refills), and self-help (such as information services). Their key findings regarding drivers of satisfaction with self-service technologies were:

- *Better than alternative*: Self-service provided a benefit over traditional service, such as saving time, easy to use, available any time, saving money, available anywhere, and avoiding service personnel.
- *Did its job*: Satisfaction emerged from an element of surprise that the technology worked as intended.

- *Solved intense need*: Transactions that include a sense of urgency involve an intense need, which is especially powerful when combined with the "always there" nature of self-service technologies.

The key drivers of dissatisfaction with self-service technologies were:

- *Technology failure*: The technology simply didn't work, resulting in problems that would affect any user.
- *Poor design*: Problem with design that would affect some, but not necessarily all users (e.g., difficult enrollment or login procedure).
- *Process failure*: Problem with process after successful completion of the initial client-technology interaction (e.g., problem with billing or delivery).
- *Client-driven failure*: Problem in which clients believe they bear some responsibility for the failure (e.g., forgot password).

Based on this research (Bitner, Ostrom, & Meuter, 2002; Ostrom, Bitner, & Meuter, 2002), Bitner et al. provided six key points for successful implementation of self-service technologies:

- *Be very clear on the purpose of the self-service technology*: Is it primarily for cost reduction, customer satisfaction, competitive positioning, or some combination?
- *Maintain a customer focus*: Understand customer needs through interviews, surveys, and focus groups. Design for usability, and test to ensure a usable design.
- *Actively promote the use of self-service technologies*: Make customers aware of the existence of the self-service option.
- *Prevent and manage failures*: Failure of technologies and service are the primary reasons why customers stop using SSTs. Because it is not possible to prevent all failures, it is important to plan for service recovery.
- *Offer choices*: Customers expect to be able to interact with service providers using whatever method they prefer. Do not force usage of self-service. Especially do not punish customers who try a self-service technology by providing no option to communicate with a live person. If possible, provide incentives for use of self-service.
- *Be prepared for constant updating and continuous improvement*: An initial self-service design will have room for improvement, perhaps due to improvements in technology, identification of additional opportunities for self-

service, or changes in leading practice in the design of user interfaces.

These points from service design, which has its roots in market research, echo key points from UCD and design thinking. Its high-level design principles include having a clear understanding of user and service provider needs, maintaining a customer/client/user focus, and iterative design.

3 EVALUATING USABILITY AND UX

3.1 Major Usability and UX Evaluation Methods

Evaluation methods for usability and UX fall into two broad classes: (1) user-based (Dumas, 2003); and (2) inspection (Cockton, Lavery, & Woolrych, 2003). Most of this section of the chapter will provide extended discussion of two of the most popular user-based evaluation methods, *usability testing* and *standardized questionnaires*. Following is a brief review of other user-based and inspection methods.

Common user-based methods include:

- *Card sorting*: Card sorting is one of the more popular ways to create and test a taxonomy or navigation structure. In an open card sort, target users organize a representative set of items into groups and then label the groups, either using physical cards or card-sorting software. In a closed sort, groups are predefined. The data are analyzed with advanced statistical methods to inform appropriate organization of content.
- *Eye tracking*: Eye tracking hardware allows UX researchers to acquire data about users' visual interactions with a display. It is most effective when visuals are combined with other metrics or data the eye-tracking software produces, producing heatmaps, focus maps, and gaze path plots. According to the 2018 UXPA salary survey, only about 10% of practitioners reporting using this method (Sauro, 2018f).
- *Surveys*: In UX evaluation, surveys can be used to capture retrospective ratings of a user's experience with a product or website. Survey respondents might be asked to consider their entire experience or to focus on a recent interaction. These

types of surveys often include one or more of the standardized questionnaires discussed later in this section (Grier, Bangor, Kortum, & Peres, 2013; Kortum & Bangor, 2013).

- *Analytics*: Software systems can capture information from which it is possible to track users' paths (Sauro, 2015a). For transactional systems, it is often possible to determine the steps at which users abandon the system, seek human assistance, or complete a task. Using analytics to discover and resolve the pain points in a system can improve numerous metrics, such as conversion rates and customer attitude toward an enterprise.
- *A/B testing:* The usability lab is great for simulating experience and testing more prominent design changes, but it can be hard to collect a large enough sample to analyze subtler alterations. In an A/B test, website visitors are randomly assigning to one of two design options (A and B). A/B testing is especially important for high-traffic websites where even a modest difference of 1 percentage point on product purchases can translate into millions of dollars of profit or loss over the course of a year.

Inspection methods emerged in the late 1980s as "discount" alternatives to usability testing (Nielsen, 1989) because they did not require the involvement of users, making them usually shorter in duration and less expensive than usability testing. Common inspection methods include:

- *Heuristic evaluation:* In a heuristic evaluation, an expert in usability principles reviews an interface against a small set of broad principles called heuristics, usually derived from analyzing the root causes of problems uncovered in usability tests. Evaluators then inspect an interface to determine how well it conforms to these heuristics and identify shortcomings. The most famous set of heuristics was derived by Nielsen and Molich (1990), but there are others. For heuristic evaluations, if you must make a tradeoff between having a single evaluator spend a lot of time examining an interface vs. having more examiners spend less time each examining an interface, choose the latter option (Dumas, Sorce, & Virzi, 1995; Virzi, 1997).
- *Guideline review:* When specific guidelines are available for a given context of use, evaluators can use them to check for inconsistencies with the guidelines. Guidelines differ from

heuristics in their specificity. For example, Smith and Mosier (1986) prepared design guidelines for the US Air Force which had six sections (e.g., data entry, data display, user guidance) with a total of 944 guidelines.

- *Keystroke Level Modeling (KLM):* Card et al. (1983) developed Keystroke Level Modeling (KLM), drawing upon research in cognitive psychology and their own empirical studies. With KLM, an evaluator can estimate how long it will take a skilled user to complete a step in a task using only a few standard operations (pointing, clicking, typing, and thinking). KLM has been shown to estimate error-free task time to within 10–30% of actual times.
- *Cognitive walkthrough:* The cognitive walkthrough is a usability inspection method similar to a heuristic evaluation (which was developed around the same time). The cognitive walkthrough has more emphasis on task scenarios than the heuristic evaluation. The cognitive walkthrough's emphasis is on learnability for first time or occasional users. As part of conducting a cognitive walkthrough, an evaluator must first identify the users' goals and how they would attempt to accomplish them in the interface. An expert in usability principles then meticulously goes through each step, identifying problems users might encounter as they learn to use the interface. In Spencer's (2000) Streamlined Cognitive Walkthrough the evaluator determines at each step (1) will users know what to do? and (2) if they do the right thing. how will they know?
- *Practical Usability Rating by Experts (PURE):* In the PURE method, an extension of the cognitive walkthrough, multiple evaluators (ideally, experts in HCI principles and the product domain) decompose user tasks and rate task difficulty on a three-point scale (Rohrer et al., 2016). The ratings are used to generate PURE scores at the task and product levels. The output is both an executive-friendly dashboard and a diagnostic set of issues uncovered as part of the task review.

3.2 What Is Usability Testing?

Usability testing is an essential skill for usability and UX practitioners. It is by no means the *only* skill in which they must have proficiency (Uldall-Espersen, Frøkjær, & Hornbæk, 2008), but it is an important one. Surveys of experienced usability practitioners consistently reveal the importance of usability testing

(Lindgaard, 2014; Mao, Vredenburg, Smith, & Carey, 2005; Sauro, 2018f; Vredenburg, Isensee, & Righi, 2002).

Imagine the two following scenarios:

> *Scenario 1* Mr. Smith is sitting next to Mr. Jones, watching him work with a high-fidelity prototype of a Web browser for a smart phone. Mr. Jones is the third person that Mr. Smith has watched performing these tasks with this version of the prototype. Mr. Smith is not constantly reminding Mr. Jones to talk while he works but is counting on his proximity to Mr. Jones to encourage verbal expressions when Mr. Jones encounters any difficulty in accomplishing his current task. Mr. Smith takes written notes whenever this happens and also takes notes whenever he observes Mr. Jones faltering in his use of the application (e.g., has trouble finding a desired function). Later that day he will use his notes to develop problem reports and, in consultation with the development team, will work on recommendations for product changes that should eliminate or reduce the impact of the reported problems. When a new version of the prototype is ready, he will resume testing.
>
> *Scenario 2* Dr. White is watching Mr. Adams work with a new version of a word-processing application. Mr. Adams is working alone in a test cell that looks almost exactly like an office, except for the large mirror on one wall and the two video cameras overhead. He has access to a telephone and a number to call if he encounters a difficulty that he cannot overcome. If he places such a call, Dr. White will answer and provide help modeled on the types of help provided at the company's call centers. Dr. White can see Mr. Adams through the one-way glass as she coordinates the test. She has one assistant working the video cameras for maximum effectiveness and another who is taking time-stamped notes on a computer (coordinated with the video time stamps) as different members of the team notice and describe different aspects of Mr. Adams's task performance. Software monitors Mr. Adams's computer, recording all keystrokes and mouse movements. Later that day, Dr. White and her associates will put together a summary of the task performance measurements for the tested version of the application, noting where the performance measurements do not meet the test criteria. They will also create a prioritized list of problems and recommendations, along with video clips that illustrate key problems, for presentation to the development team at their weekly status meeting.

Both of these scenarios provide examples of usability testing. In scenario 1 the emphasis is completely on usability problem discovery and resolution (formative/diagnostic/qualitative evaluation). In scenario 2 the primary emphasis is on task performance measurement (summative/measurement-focused/quantitative evaluation), but there is also an attempt to record and present usability problems to the product developers. Dr. White's team knows that they cannot determine if they've met the usability performance goals by examining a list of problems, but they also know that they cannot provide appropriate guidance to product development if they present only a list of global task measurements. The problems observed in the use of an application provide important clues for redesigning the product (Chapanis, 1981; Norman, 1983). Furthermore, as J. Karat (1997, p. 693) observed: "The identification of usability problems in a prototype user interface (UI) is not the end goal of any evaluation. The end goal is a redesigned system that meets the usability objectives set for the system such that users are able to achieve their goals and are satisfied with the product."

These scenarios also illustrate the defining properties of a usability test. During a usability test, one or more observers watch one or more participants perform specified tasks with the product in a specified test environment (compare this with the ISO/ANSI definition of usability presented earlier in this chapter). This observation of actual use in a controlled setting is what makes usability testing different from other user-centered design (UCD) methods or market research (Dumas & Salzman, 2006). In interviews (including the group interview known as a focus group), participants do not perform worklike tasks. Usability inspection methods (such as expert evaluations and heuristic evaluations) also do not include the observation of users or potential users performing worklike tasks. The same is true of techniques such as surveys and card sorting. Field studies (including contextual inquiry) can involve the observation of users performing work-related tasks in target environments but restrict the control that practitioners have over the target tasks and environments. Note that this is not necessarily a bad thing, but it is a defining difference between usability testing and field (ethnographic) studies.

This definition of usability testing permits a wide range of variation in technique (Wildman, 1995). Usability tests can be very informal (as in scenario 1) or very formal (as in scenario 2). The observer might sit next to the participant, watch through a one-way glass, watch the on-screen behavior of a participant who is performing specified tasks at a location halfway around the world, or set up a remote unmoderated usability test in which participants complete tasks online at any time. Usability tests can be think-aloud (TA) tests, in which observers train participants to talk about what they're doing at each step of task completion and prompt participants to continue talking if they stop. Observers might watch one participant at a time or might watch participants work in pairs. Practitioners can apply usability testing to the evaluation of low-fidelity prototypes (MacKenzie &

Read, 2007), high-fidelity prototypes, mixed-fidelity prototypes (McCurdy et al., 2006), Wizard of Oz (WoZ) prototypes (Dow et al., 2005; Kelley, 1985), products under development, predecessor products, or competitive products.

3.2.1 Where Did Usability Testing Come From?

The roots of usability testing lie firmly in the experimental methods of psychology (in particular, cognitive and applied psychology) and human factors engineering (Dumas & Salzman, 2006) with strong ties to the concept of iterative design. In a traditional experiment, the experimenter draws up a careful plan of study that includes the exact number of participants that the experimenter will expose to the different experimental treatments. The participants are members of the population to which the experimenter wants to generalize the results. The experimenter provides instructions and debriefs the participant, but at no time during a traditional experimental session does the experimenter interact with the participant (unless this interaction is part of the experimental treatment).

The more formative (diagnostic, focused on problem discovery) the focus of a usability test, the less it is like a traditional experiment (although the requirements for sampling from a legitimate population of users, tasks, and environments still apply). Conversely, the more summative (focused on measurement) a usability test is, the more it should resemble the mechanics of a traditional experiment. Many of the principles of psychological experimentation that exist to protect experimenters from threats to reliability and validity (e.g., the control of demand characteristics, the Hawthorne effect) carry over into usability testing (Holleran, 1991; Macefield, 2007; Wenger & Spyridakis, 1989).

3.2.2 Is Usability Testing Effective?

The widespread use of usability testing is evidence that practitioners believe that usability testing is effective. Unfortunately, there are fields in which practitioners' belief in the effectiveness of their methods does not appear to be warranted by those outside the field (e.g., the use of projective techniques such as the Rorschach test in psychotherapy; Lilienfeld. Wood, & Garb, 2000). In our own field, papers published since 1998 have questioned the reliability of usability problem discovery (Hertzum & Jacobsen, 2003; Hertzum, Molich, & Jacobsen, 2014; Kessner et al., 2001; Molich et al., 1998; Molich & Dumas, 2008; Molich, Ede, Kaasgaard, & Karyukin).

The common finding in these studies has been that observers (either individually or in teams across usability laboratories) who evaluated the same product produced markedly different sets of discovered problems. Molich et al. (1998) had four independent usability laboratories carry out inexpensive usability tests of a software application for new users. The four teams reported 141 different

problems, with only one problem common among all four teams. Molich et al. (1998) attributed this inconsistency to variability in the approaches taken by the teams (task scenarios, level of problem reporting). Kessner et al. (2001) had six professional usability teams independently test an early prototype of a dialog box. None of the problems were detected by every team, and 18 problems were described by one team only. Molich et al. (2004) assessed the consistency of usability testing across nine independent organizations that evaluated the same website. They documented considerable variability in methodologies, resources applied, and problems reported. The total number of reported problems was 310, with only 2 problems reported by 6 or more organizations, and 232 problems (61%) uniquely reported. The fourth comparative usability evaluation (CUE-4; Molich & Dumas, 2008) had a similar method and similar outcomes. "Our main conclusion is that our simple assumption that we are all doing the same and getting the same results in a usability test is plainly wrong" (Molich et al., 2004, p. 65).

This is important and disturbing research, but there is a clear need for more research in this area. A particularly important goal of future research should be to reconcile these studies with the documented reality of usability improvement achieved through iterative application of usability testing. For example, a limitation of research that stops with the comparison of problem lists is that it is not possible to assess the magnitude of the usability improvement (if any) that would result from product redesigns based on design recommendations derived from the problem lists (Hornbæk, 2010; Wixon, 2003). When comparing problem lists from many labs, one aberrant set of results can have an extreme effect on measurements of consistency across labs, and the more labs that are involved, the more likely this is to happen.

In the case of CUE-4 (Molich & Dumas, 2008), 17 professional usability teams evaluated the same website, with 9 teams conducting usability tests (5–15 participants per test) and 8 teams using expert review (1–2 reviewers per team). With one exception, the usability test teams used different sets of tasks for their evaluations. Across the 17 teams, there were 76 usability test participants and 10 expert reviewers, for a total of 86 individual experiences with the website. Using the binomial probability formula, it is possible to estimate the percentage of problems discovered with this sample size for problems of different likelihoods of discovery (Sauro & Lewis, 2016). For individual problems that would affect 10% of participants, the likelihood of having the problem turn up *at least once* in this study is about 99.99%, making their discovery virtually certain. For problems with a 1% probability of occurrence, the likelihood of discovery (at least once) with a sample size of 86 is about 58%, better than even odds. Even problems with probabilities of occurrence as low as 0.1% had about an 8% likelihood of discovery. It is not possible to know how many specific problems were available for discovery as a function of their probabilities of occurrence, but it seems reasonable that a mature website would have eliminated most high-probability

problems, leaving a mass of less probable (hard-to-discover) problems, leading to little overlap in problem discovery across the teams. As Molich and Dumas (2008, p. 270) concluded, "The limited overlap could be interpreted as a sign that some of the teams … had conducted a poor evaluation. Our interpretation, however, is that the usability problem space is so huge that it inevitably leads to some instances of limited overlap." Furthermore, difficulties in matching problem descriptions can lead to an appearance of greater underlap than occurs when observers have an opportunity to discuss problem matching (Hornbæk, 2010; Hornbæk & Frøkjær, 2008a, 2008b).

Hertzum et al. (2014) published results from CUE 9, in which 19 experienced usability professionals analyzed videos of test sessions with five users who worked on reservation tasks at a truck rental website. Nine professionals analyzed moderated sessions; ten analyzed unmoderated sessions. The patterns of problem discovery across professionals were similar for moderated and unmoderated testing, and were similar to previous related research in that agreement among the usability professionals was not perfect.

The interpretation of the results of these studies (Kessner et al., 2001; Molich et al., 1998; Molich et al., 2004; Molich & Dumas, 2008) as indicative of a lack of reliability (e.g., Law et al., 2005) stands in stark contrast to the published studies in which iterative usability tests (sometimes in combination with other UCD methods) have led to significantly improved products (Al-Awar et al., 1981; Bailey, 1993; Bailey, Allan, & Raiello 1992; Gould et al., 1987; Kelley, 1984; Kennedy, 1982; Lewis, 1982, 1996; Ruthford & Ramey, 2000). For example, in a paper describing their experiences in product development, Marshall et al. (1990, p. 243) stated: "Human factors work can be reliable—different human factors engineers, using different human factors techniques at different stages of a product's development, identified many of the same potential usability defects." Published cost–benefit analyses (Bias & Mayhew, 1994) have demonstrated the value of usability engineering processes that include usability testing, with cost–benefit ratios ranging from 1:2 for smaller projects to 1:100 for larger projects (C. Karat, 1997).

Most of the papers that describe the success of iterative usability testing are case studies (such as Høegh & Jensen, 2008; Marshall, Brendan, & Prail, 1990―for an adaptation of usability testing to an Agile framewor, see Sy, 2007; Illmensee & Muff, 2009), but a few have described designed experiments. Bailey et al. (1992) compared two user interfaces derived from the same base interface: one modified via heuristic evaluation and the other modified via iterative usability testing (three iterations, five participants per iteration). They conducted this experiment with two interfaces, one character based and the other a graphical user interface (GUI), with the same basic outcomes. The number of changes indicated by usability testing was much smaller than the number indicated by heuristic evaluation, but user performance was the same with both final versions of the

interface. All designs after the first iteration produced faster performance and, for the character-based interface, were preferred to the original design. The time to complete the performance testing was about the same as that required for the completion of multireviewer heuristic evaluations.

Bailey (1993) provided additional experimental evidence that iterative design based on usability tests leads to measurable improvements in the usability of an application. In the experiment, he studied the designs of eight designers, four with at least four years of professional experience in interface design and four with at least five years of professional experience in computer programming. All designers used a prototyping tool to create a recipes application (eight applications in all). In the first wave of testing, Bailey videotaped participants performing tasks with the prototypes, three different participants per prototype. Each designer reviewed the videotapes of the people using his or her prototype and used the observations to redesign his or her application. This process continued until each designer indicated that it was not possible to improve his or her application. All designers stopped after three to five iterations. Comparison of the first and last iterations indicated significant improvement in measurements such as number of tasks completed, task completion times, and repeated serious errors.

In conclusion, the results of the studies of Molich et al. (1998, 2004; Molich & Dumas, 2008) and similar studies show that usability practitioners must conduct their usability tests as carefully as possible, document their methods completely, and show proper caution when interpreting their results. Agreement isn't necessarily the key goal of iterative design and small sample assessment—the ultimate goal is improved usability and a better user experience. The limitations of usability testing make it insufficient for certain testing goals, such as quality assurance of safety-critical systems (Thimbleby, 2007). It can be difficult to assess complex systems with complex goals and tasks (Howard, 2008; Howard & Howard, 2009; Redish, 2007). On the other hand, as Landauer stated (1997, p. 204): "There is ample evidence that expanded task analysis and formative evaluation can, and almost always do, bring substantial improvements in the effectiveness and desirability of systems."

3.3 Goals of Usability Testing

The fundamental goal of usability testing is to help developers produce more usable products. The two conceptions of usability testing (formative and summative) lead to differences in the specification of goals in much the same way that they contribute to differences in fundamental definitions of usability (diagnostic problem discovery and measurement). Rubin (1994, p. 26) expressed the formative goal as follows: "The overall goal of usability testing is to identify and rectify usability deficiencies existing in computer-based and electronic equipment and their accompanying support materials prior to release." Dumas and

Redish (1999, p. 11) provided a more summative goal: "A key component of usability engineering is setting specific, quantitative, usability goals for the product early in the process and then designing to meet those goals."

These goals are not in direct conflict, but they do suggest different focuses that can lead to differences in practice. For example, a focus on measurement typically leads to more formal testing (less interaction between observers and participants), whereas a focus on problem discovery typically leads to less formal testing (more interaction between observers and participants). In addition to the distinction between diagnostic problem discovery and measurement tests, there are two common types of measurement tests: comparison against objectives and comparison of products.

3.3.1 Problem Discovery Test

The primary activity in diagnostic problem discovery tests is the discovery, prioritization, and resolution of usability problems. The number of participants in each iteration of testing should be fairly small, but the overall test plan should be for multiple iterations, each with some variation in participants and tasks. When the focus is on problem discovery and resolution, the assumption is that more global measures of user performance and satisfaction will take care of themselves (Chapanis, 1981). The measurements associated with problem discovery tests are focused on prioritizing problems and include frequency of occurrence in the test, likelihood of occurrence during normal usage (taking into account the anticipated usage of the part of the product in which the problem occurred), and magnitude of impact on the participants who experienced the problem. Because the focus is not on precise measurement of the performance or attitudes of participants, problem discovery studies tend to be informal, with a considerable amount of interaction between observers and participants. Some typical stopping rules for iterations are a preplanned number of iterations or a specific problem discovery goal, such as "Identify 90% of the problems available for discovery for these types of participants, this set of tasks, and these conditions of use." As Lindgaard (2006, p. 1069) pointed out:

> It is impossible to know whether all usability problems have been identified in a particular test or type of evaluation unless testing is repeated until it reaches an asymptote, a point at which no new problems emerge in a test. Asymptotic testing is not, and should not be, done in practice; it is as unfeasible as it is irrelevant in a work context.

3.3.2 Measurement Test Type I: Comparison against Quantitative Objectives

Studies that have a primary focus of comparison against quantitative objectives include two fundamental activities (Jokela et al., 2006; Sauro, 2018a). The first is the development of the usability objectives. The second is iterative testing to determine if the product has met the objectives. A third activity (which can take place during iterative testing) is the enumeration and description of usability problems, but this activity is secondary to the collection of precise measurements.

The first step in developing quantitative usability objectives is to determine the appropriate variables to measure. As part of the work done for the European MUSiC (Measuring the Usability of Systems in Context) project, Rengger (1991) produced a list of potential usability measurements based on 87 papers out of a survey of 500 papers. He excluded purely diagnostic studies and also excluded papers if they did not provide measurements for the combined performance of a user and a system. He categorized the measurements into four classes:

- *Class 1:* goal achievement indicators (such as success rate and accuracy)
- *Class 2:* work rate indicators (such as speed and efficiency)
- *Class 3:* operability indicators (such as error rate and function usage)
- *Class 4:* knowledge acquisition indicators (such as learnability and learning rate)

In a later discussion of the MUSiC measures, Macleod et al. (1997) described measures of effectiveness (the level of correctness and completeness of goal achievement in context) and efficiency (effectiveness related to cost of performance, typically the effectiveness measure divided by task completion time). Optional measures were of productive time and unproductive time, with unproductive time consisting of help actions, search actions, and snag (negation, canceled, or rejected) actions.

Macleod et al.'s (1997) description of the measures of effectiveness and efficiency seem to have influenced the objectives expressed in ISO 9241-11 (ISO, 1998, p. iv):

> The objective of designing and evaluating visual display terminals for usability is to enable users to achieve goals and meet needs in a particular context of use. ISO 9241-11 explains the benefits of measuring usability in terms of user performance and satisfaction. These are measured by the extent to which the intended goals of use are achieved, the resources that have to be expended to achieve the intended goals, and the extent to which the user finds the use of the product acceptable.

In practice, and as recommended by ANSI (2001), the fundamental global measurements for usability tasks are successful task completion rates (for a measure of effectiveness), mean task completion times (for a measure of efficiency—either the arithmetic mean or, as recently suggested by Sauro and Lewis (2010), the geometric mean), and mean participant satisfaction ratings (collected either on a task-by-task basis or at the end of a test session; see Section 3.9 for more information on measuring participant satisfaction and other elements of user experience). There are many other measurements that practitioners could consider (Dumas & Redish, 1999; Nielsen, 1997), including but not limited to (1) the number of tasks completed within a specified time limit; (2) the number of wrong menu choices; (3) the number of user errors; and (4) the number of repeated errors (same user committing the same error more than once).

After determining the appropriate measurements, the next step is to set the goals. Ideally, the goals should have an objective basis and shared acceptance across the various stakeholders, such as marketing, development, and test groups (Lewis, 1982). The best objective basis for measurement goals is data from previous usability studies of predecessor or competitive products. For maximum generalizability, the historical data should come from studies of similar types of participants completing the same tasks under the same conditions (Chapanis, 1988). If this information is not available, an alternative is for the test designer to recommend objective goals and to negotiate with the other stakeholders to arrive at a set of shared goals.

According to Rosenbaum (1989, p. 211):

> Defining usability objectives (and standards) isn't easy, especially when you're beginning a usability program. However, you're not restricted to the first objective you set. The important thing is to establish some specific objectives immediately, so that you can measure improvement. If the objectives turn out to be unrealistic or inappropriate, you can revise them.

Such revisions, however, should take place only in the early stages of gaining experience and taking initial measurements with a product. It is important not to change reasonable goals to accommodate an unusable product.

When setting usability goals, it is usually better to set goals that refer to an average (mean) of a measurement than to a percentile. For example, set an objective such as "The mean time to complete task 1 will be less than 5 minutes" rather than "95% of participants will complete task 1 in less than 10 minutes." The statistical reason for this is that sample means drawn from a continuous distribution are less variable than sample medians (the 50th percentile of a sample), and measurements made away from the center of a distribution (e.g.,

measurements made to attempt to characterize the value of the 95th percentile) are even more variable (Blalock, 1972). Cordes (1993) conducted a Monte Carlo study comparing means and medians as measurements of central tendency for time-on-task scores and determined that the mean should be the preferred metric for usability studies (unless there is missing data due to participants failing to complete tasks, in which case the mean from the study will underestimate the population mean). Sauro and Lewis (2010) have recommended reporting the geometric mean rather than the median for task times, and Rummel (2014, 2017) has published a number of advanced methods for characterizing task times.

A practical reason to avoid percentile goals is that the goal can imply a sample size requirement that is unnecessarily large. For example, you cannot measure accurately at the 95th percentile unless there are at least 20 measurements (in fact, there must be many more than 20 measurements for accurate measurement). An exception to this is the specification of successful task completions (or any other measurement that is based on counting events), which necessarily requires a percentile goal, usually set at or near 100% (unless there are historical data that indicate an acceptable lower level for a specific test). If 10 out of 10 participants complete a task successfully, the observed completion rate is 100%, but a 90% exact binomial confidence interval for this result ranges from 74% to 100%. In other words, even perfect performance for 10 participants with this type of measure leaves open the possibility (with 90% confidence) that the true completion rate could be as low as 75%. See Sauro and Lewis (2016) for more information on computing and using this information in usability tests.

After the usability goals have been established, the next step is to collect data to determine if the product has met its goals. Representative participants perform the target tasks in the specified environment as test observers record the target measurements and identify, to the extent possible within the constraints of a more formal testing protocol, details about any usability problems that occur. The usability team conducting the test provides information about goal achievement and prioritized problems to the development team, and a decision is made regarding whether or not there is sufficient evidence that the product has met its objectives. The ideal stopping rule for measurement-based iterations is to continue testing until the product has met its goals.

When there are only a few goals, it is reasonable to expect to achieve all of them. When there are many goals (e.g., 5 objectives per task multiplied by 10 tasks, for a total of 50 objectives), it is more difficult to determine when to declare success and to stop testing. Thus, it is sometimes necessary to specify a metaobjective of the percentage of goals to achieve.

Despite the reluctance of some usability practitioners to conduct statistical tests to quantitatively assess the strength of the available evidence regarding whether or not a product has achieved a particular goal, the leading practice is to

conduct such tests. The best approach is to conduct multiple *t*-tests or nonparametric analogs of *t*-tests (Lewis, 1993) because this gives practitioners the level of detail that they require. There is a well-known prohibition against doing this because it can lead investigators to mistakenly accept as real some differences that are due to chance (technically, alpha (α) inflation). Note that standard approaches to controlling alpha inflation, such as analysis of variance (ANOVA) and multiple comparisons methods (e.g., Bonferroni or Benjamini-Hochberg—see Sauro & Lewis, 2016 for details), are appropriate when comparing sets of means, but not when comparing means with benchmarks.

Furthermore, if the required level of information is at the level of individual *t*-tests, it is an appropriate method (Abelson, 1995). The practice of avoiding alpha inflation is a concern more related to scientific hypothesis testing than to usability testing (Wickens, 1998), although usability practitioners should be aware of its existence and take it into account when interpreting their statistical results. For example, if you compare two products by conducting 50 *t*-tests with alpha set to 0.10, and only 5 (10%) of the *t*-tests are significant (have a *p*-value below 0.10), you should question whether or not to use those results as evidence of the superiority of one product over the other. On the other hand, if substantially more than 5 of the *t*-tests are significant, you can be more confident that the differences indicated are real.

In addition to (or as an alternative to) conducting multiple *t*-tests, practitioners should compute confidence intervals for their measurements. This applies to the measurements made for the purpose of establishing test criteria (such as measurements made on predecessor versions of the target product or competitive products) and to the measurements made when testing the product under development. See Sauro and Lewis (2016) for more details.

3.3.3 Measurement Test Type II: Comparison of Products

The second type of measurement test is to conduct usability tests for the purpose of direct comparison of one product with another. As long as there is only one measurement that decision makers plan to consider, a standard *t*-test or ANOVA with multiple comparisons (ideally, in combination with the computation of confidence intervals) will suffice for the purpose of determining whether one product is superior. When there are multiple dependent variables (e.g., completion time and satisfaction) and no compelling need to combine them into a single metric, you can run multiple *t*-tests or ANOVAs on each dependent variable.

If decision makers want to combine multiple dependent measures into an overall metric to determine which product has the best overall usability, standard multivariate statistical procedures (such as multivariate analysis of variance

(MANOVA) or discriminant analysis) are not often helpful. The statistical reason for this is that multivariate statistical procedures depend on the computation of centroids (a weighted average of multiple dependent measures) using a least-squares linear model that maximizes the difference between the centroids of the two products (Cliff, 1987). If the directions of the measurements are inconsistent (e.g., a high task completion rate is desirable, but a high mean task completion time is not), the resulting centroids are uninterpretable for the purpose of usability comparison. In some cases it is possible to recompute variables so they have consistent directions (e.g., recomputing task completion rates as task failure rates). If this is not possible, another approach is to convert measurements to ranks (Lewis, 1991a) or standardized (*Z*) scores (Sauro & Kindlund, 2005) for the purpose of principled combination of different types of measurements. After this combination, it is reasonable to conduct analyses with ANOVA or *t*-tests for standardized scores, or with their nonparametric analogs for ranks.

To help consumers compare the usability of different products, ANSI (2001) has published the Common Industry Format (CIF) for usability test reports. Originally developed at the National Institute of Standards and Technology (NIST), this test format requires measurement of effectiveness (accuracy and completeness—completion rates, errors, assists), efficiency (resources expended in relation to accuracy and completeness—task completion time), and satisfaction (freedom from discomfort, positive attitude toward use of the product—using any of a number of standardized satisfaction questionnaires). It also requires a complete description of participants and tasks.

Morse (2000) reviewed a NIST project conducted to pilot test the CIF. The purpose of the CIF is to make it easier for purchasers to compare the usability of different products. The pilot study ran into problems, such as inability to find a suitable software product for both supplier and consumer, reluctance to share information, and uncertainty about how to design a good usability study. To date, there has been little if any use (at least, no published use) of the CIF for its intended purpose.

3.4 Variations on a Theme: Other Types of Usability Tests

3.4.1 Think Aloud (TA)

In a standard, formal usability test, test participants perform tasks without necessarily speaking as they work. The defining characteristic of a TA study is the instruction to participants to talk about what they are doing as they do it (in other words, to produce verbal reports). If participants stop talking (as commonly

happens when they become very engaged in a task), they are prompted to resume talking.

The most common theoretical justification for the use of TA is from the work in cognitive psychology (specifically, human problem solving) of Ericsson and Simon (1980). Responding to a review by Nisbett and Wilson (1977) that described various ways in which verbal reports were unreliable, Ericsson and Simon provided evidence that certain kinds of verbal reports could produce reliable data. They stated that reliable verbalizations are those that participants produce during task performance that do not require additional cognitive processing beyond the processing required for task performance and verbalization, either when the verbal expression is already in the participant's attention in verbal form (Level 1) or in the participant's attention in nonverbal form (Level 2). Level 3 verbalizations, the expression of information not currently in the participant's attention such as descriptions of reasons, explanations, or feelings, can affect thought processes and behavior, so Ericsson and Simon excluded them from their conception of TA (Hertzum & Holmegaard, 2013).

TA is not feasible when testing systems that include speech recognition (Lewis, 2008, 2011b). For usability testing of other systems, the use of TA is fairly common. Dumas (2003) encouraged the use of TA because (1) TA tests are more productive for finding usability problems (van den Haak & de Jong, 2003; Virzi et al., 1993), and (2) thinking aloud does not affect user ratings or performance (Bowers & Snyder, 1990; Ohnemus & Biers, 1993; Olmsted-Hawala, Murphy, Hawala, & Ashenfelter, 2010). There is some evidence in support of these statements, but the evidence is mixed.

Earlier prohibitions against the use of TA in measurement-based tests assumed that thinking aloud would cause slower task performance. Bowers and Snyder (1990), however, found no measurable task performance or preference differences between a test group that thought aloud and one that did not. Surprisingly, there are some experiments in which the investigators reported better task performance when participants were thinking aloud. Berry and Broadbent (1990) provided evidence that the process of thinking aloud invoked cognitive processes that improved rather than degraded performance, but only if people were given (1) verbal instructions on how to perform the task, and (2) the requirement to justify each action aloud. Wright and Converse (1992) compared silent with TA usability testing protocols. The results indicated that the TA group committed fewer errors and completed tasks faster than the silent group, and the difference between the groups increased as a function of task difficulty.

Regarding the theoretical justification for and typical practice of TA, Boren and Ramey (2000) noted that TA practice in usability testing often does not conform to the theoretical basis most often cited for it (Ericsson & Simon, 1980). "If practitioners do not uniformly apply the same techniques in conducting

thinking-aloud protocols, it becomes difficult to compare results between studies" (Boren & Ramey, 2000, p. 261). In a review of publications of TA tests and field observations of practitioners running TA tests, they reported inconsistency in explanations to participants about how to TA, practice periods, styles of reminding participants to TA, prompting intervals, and styles of intervention. They suggest that, rather than basing current practice on Ericsson and Simon, a better basis would be speech communication theory, with clearly defined communicative roles for the participant (in the role of domain expert or valued customer, making the participant the primary speaker) and the usability practitioner (the learner or listener, thus a secondary speaker).

Based on this alternative perspective for the justification of TA, Boren and Ramey (2000) provided guidance for many situations that are not relevant in a cognitive psychology experiment but are in usability tests. For example, they recommend that usability practitioners running a TA test should continually use acknowledgment tokens that do not take speakership away from the participant, such as "mm hm?" and "uh-huh?" (with the interrogative intonation) to encourage the participant to keep talking. In normal communication, silence (as recommended by the Ericsson and Simon protocols) is not a nonresponse—the speaker interprets it in a primarily negative way as indicating aloofness or condescension. They avoided providing precise statements about how frequently to provide acknowledgments or somewhat more explicit reminders (such as "And now…?") because the best cues come from the participants. Practitioners need to be sensitive to these cues as they run the test.

Krahmer and Ummelen (2004) conducted an exploratory comparison of the standard Ericsson and Simon versus the Boren and Ramey speech-communication TA procedures (10 participants per condition). They found that the outcomes were similar for both procedures, with participants in both conditions saying about the same number of words, uncovering essentially the same navigation problems, and providing about equal evaluations of the quality of the website they used. The main difference was that moderators in the speech-communication condition made, as expected, more interventions and, perhaps as a consequence, the participants seemed less lost and completed more tasks.

Hertzum et al. (2009) compared silent task completion with strict and more relaxed TA, supplemented with eye tracking and assessment of mental workload. Strict TA, other than requiring more time for task completion, led to similar results as the silent condition. Relaxed TA affected participant behavior in multiple ways. Relative to silence, the TA method did not affect successful task completion rates, which tended to be high in the study. In the relaxed TA condition, participants spent more time in general distributed visual behavior, issued more navigation commands, and experienced higher mental workload.

Olmsted-Hawala et al. (2010) used a double-blind procedure to investigate the effect of different TA procedures on successful task completion, task completion times, and satisfaction. Their experimental conditions were the traditional TA, speech-communication TA, a less restrictive coaching protocol in which moderators could freely probe participants (i.e., active intervention), and silence (no TA at all), with 20 participants per condition. The outcomes were similar for silence, standard, and speech-communication TA procedures. Participants in the coaching condition successfully completed significantly more tasks and had higher satisfaction ratings. Their results for speech-communication differed from those reported by Krahmer and Umullen (2004): "since the test administrator in the Krahmer and Umullen study offered assistance and encouragement to the test subject during the session, we think their speech-communication protocol is more akin to the coaching condition in our study" (Olmsted-Hawala et al., 2010, p. 2387).

McDonald, McGarry, and Willis (2013) compared the standard TA procedure ("say out loud everything you say to yourself silently") with a method in which participants were instructed to explain their actions. Participants attempted relatively easy and relatively difficult navigation tasks with a web-based encyclopedia. They found an interaction between method and task difficulty such that task success was about the same with both methods for easier tasks, but better for difficult tasks when explaining their actions. Task difficulty affected completion time (as expected), but the type of method did not. Consistent with their instruction, participants in the explanation condition provided more explanatory verbalizations, but in general all participants provided a high proportion of procedural descriptions.

Alhadreti and Mayhew (2017) collected data from 60 participants attempting tasks on a library website, 20 each in different TA conditions (standard, speech-communication, and active-intervention). Task completion rates, perceived usability (SUS), and overall problem identification were about the same for all three conditions. Other measures were similar for standard and speech-communication, but the active-intervention condition led to slower task completion, more mouse clicks, more pages browsed, a greater sense of distraction, and required more time to complete the study.

Hertzum and Holmegaard (2013) studied the effect of interruptions and time constraints on TA. Participants solved code-breaking puzzles presented on a computer. The independent variables of the study were TA condition (TA or not—between subjects), interruption (auditory, visual, both, or none—within subjects), and task timing (timed or not—within subjects). There were a number of significant outcomes (e.g., higher completion rates for untimed tasks), but focusing on the main effect of TA condition or interactions with that effect, there was a significant interaction between TA condition and type of interruption on the task completion rate, with participants in the TA group solving more tasks in the

presence of visual interruption, but taking more time to respond to interruptions. Effects associated with time constraints were independent of the TA condition. In a similar experiment, Hertzum and Holmegaard (2015) reported that TA affected the perceived time required to solve these types of code breaking puzzles, with the TA group's overestimation of the passage of time (47%) significantly less than the control group's overestimation (94%).

In 2012, McDonald et al. conducted an international survey to explore how much variation there was in how practitioners used TA. They received 207 responses from professional usability evaluators, drawn from personal contacts, usability companies, conference attendees, and special interest groups. Respondents reporting finding concurrent TA to be well suited to usability testing, but there was evidence of variation in practice with regard to think-aloud instruction, practice, interventions, and use of demonstrations. Respondents were aware of potential threats to test reliability and made attempts to mitigate them. McDonald et al. (p. 2) concluded, "The reliability considerations underpinning the classic think-aloud approach are pragmatically balanced against the need to capture useful data in the time available."

The evidence indicates that relative to silent participation, TA can affect task performance and reported satisfaction, depending on the exact TA protocol in use. If the primary purpose of the test is problem discovery, TA appears to have advantages over completely silent task completion. If the primary purpose of the test is task performance measurement, the use of TA is somewhat more complicated. As long as all the tasks in the planned comparisons were completed under the same conditions, performance comparisons should be legitimate. It is critical, however, that practitioners using TA provide a complete description of their method, including the kind and frequency of probing.

The use of TA almost certainly prevents generalization of task performance outside the TA task, but there are many other factors that make it difficult to generalize specific task performance data collected in usability studies. For example, Cordes (2001) demonstrated that participants assume that the tasks they are asked to perform in usability tests are possible (the "I know it can be done or you wouldn't have asked me to do it" bias). Manipulations that bring this assumption into doubt can have a strong effect on quantitative usability performance measures, such as increasing the percentage of participants who give up on a task. If uncontrolled, this bias makes performance measures from usability studies unlikely to be representative of real-world performance when users are uncertain as to whether the product can support the desired tasks.

The discussion above focuses on concurrent TA, with participants talking aloud as they perform tasks. An alternative approach is to use stimulated retrospective TA, in which participants perform tasks silently, and then talk as they review the video of their task performance—an approach that avoids any

influence of TA on task performance but requires at least twice as much time to complete data collection in a usability study. Bowers and Snyder (1990) reported similar task performance and subjective measures for concurrent and retrospective TA, but participants provided different types of information as a function of TA style, with participants in the concurrent condition tending to provide procedural information, and participants in the retrospective condition tending to give explanations and design statements.

Similar findings were reported by van den Haak and de Jong (2003), along with fewer successful task completions for TA relative to silent work. Karahasanovic et al. (2009), comparing concurrent and retrospective TA with a feedback collection method (FCM) in which participants respond to probes during task performance, found that all methods were intrusive with regard to completion rates and times, but the FCM was less time-consuming to analyze.

Using eye tracking to assess a participant's focus of attention, Guan et al. (2006) found the retrospective method to be valid and reliable, with a low risk of introducing fabrication, with no significant effect of task complexity. Elbabour et al. (2017) used eye tracking with two variants of retrospective think-aloud: video-cued and gaze-cued, reporting that the combination of TA with eye tracking helped evaluators detect more usability problems, especially minor navigational and comprehension problems. Seeing where they were looking at helped their participants remember details, but participants were sometimes distracted while watching their eye movements and stopped talking. Sauro (2017a) found significantly different viewing patterns when including eye tracking in a study in which participants worked briefly with 20 website home pages and thought aloud for a random half of the pages.

McDonald, Zhao, and Edwards (2013) used a dual-elicitation method in which 10 participants engaged in both standard concurrent and a type of retrospective TA in a study of a university intranet, but rather than viewing a video of their task performance, the retrospective cue was to read the task description and talk through their memory of what happened. They found considerable overlap in the content of the two TA sessions and concluded that the retrospective phase produced more verbalizations that were relevant to usability analysis, despite the occurrence of a small number of less desirable utterances (hypothesizing, rationalizing, forgetting).

Clemmensen et al. (2009) discussed the impact of cultural differences on TA. There are several ways in which cultural differences could affect testing, such as the instructions and tasks, the participant's verbalization, how the observer "reads" the participant, and the overall relationship between participant and observer. In particular, regarding studies that have Western observers and Eastern participants, they recommended that observers should allow sufficient time for

participants to pause while thinking aloud, rely less on expressions of surprise, and be sensitive to the tendency for indirect criticism.

3.4.2 Multiple Simultaneous Participants

Downey (2007) described group usability testing in which multiple observers watch a number of participants individually but simultaneously perform tasks. A key benefit of the method was obtaining data from more people over a shorter period of time. She reported that the method appeared to be most effective when tasks were relatively simple and a focused discussion followed the group's completion of the tasks.

Another way to encourage participants to talk during task completion is to have them work together (Wildman, 1995), a method sometimes called constructive interaction (Nielsen, 1993) or co-participation (Alhadreti & Mayhew, 2018). This strategy is similar to TA in its strengths and limitations, but with potentially greater ecological validity, including less participant awareness of the observer (van den Haak & de Jong, 2005; van den Haak, de Jong, & Schellens, 2006).

Hackman and Biers (1992) compared three TA methods: thinking aloud alone (Single), thinking aloud in the presence of an observer (Observer), and verbalizations occurring in a two-person team (Team). They found no significant differences in performance or subjective measures. The Team condition produced more statements of value to designers than the other two conditions, but this was probably due to the differing number of participants producing statements in the different conditions. There were three groups, with 10 participants per group for Single and Observer and 20 participants (10 two-person teams) for the Team condition. "The major result was that the team gave significantly more verbalizations of high value to designers and spent more time making high value comments. Although this can be reduced to the fact that the team spoke more overall and that there are two people talking rather than one, this finding is not trivial" (Hackman & Biers, 1992, p. 1208).

Alhadreti and Mayhew (2018) compared traditional concurrent TA with a condition similar to Hackman and Biers' (1992) two-person team condition in a between-subjects design with matched demographic characteristics, 20 participants per condition attempting tasks on a library website. They found no significant differences between the groups with regard to task performance (completion rate, completion time, number of mouse clicks, number of pages browsed) or perceived usability measured with the SUS. The groups differed significantly on ratings of their testing experience, with members of the two-person teams having a more positive experience than participants in the traditional TA condition. The team condition turned up more low-severity problems.

3.4.3 Remote Evaluation

Advances in the technology of collaborative software have made it easier to conduct remote software tests—tests in which the usability practitioner and the test participant are in different locations (Albert, Tullis, & Tedesco, 2010; Ramli & Jaafar, 2009). This can be an economical alternative to bringing one or more users into a laboratory for face-to-face user testing. A participant in a remote location can view the contents of the practitioner's screen, and in a typical system the practitioner can decide whether the participant can control the desktop. System performance is typically slower than that of a local test session.

Some of the advantages of remote testing are: (1) access to participants who would otherwise be unable to participate (international, special needs, etc.); (2) the capability for participants to work in familiar surroundings; and (3) no need for either party to install or download additional software other than generally lightweight screen sharing software like Zoom, Blue Jeans, GoToMeeting, and WebEx. Some of the disadvantages are: (1) potential uncontrolled disruptions in the participant's workplace; (2) lack of visual feedback from the participant; and (3) the possibility of compromised security if the participant takes screen captures of confidential material. Despite these disadvantages, McFadden et al. (2002) reported data that indicated that remote testing was effective at improving product designs and that the test results were comparable to the results obtained with more traditional testing.

As described above, synchronous remote usability testing has similar time constraints as laboratory-based tests (Dumas & Salzman, 2006). More fully automated asynchronous usability testing has become available which permits more rapid testing, typically with the participant receiving information about the task and responding to questions in one window while working with the product in a different window (West & Lehman, 2006). A clear disadvantage of this type of unmoderated testing is the lack of interaction between observers and participants, but Tullis et al. (2002) reported no substantial differences between unmoderated and laboratory testing for quantitative measurements or problem discovery.

West and Lehman (2006) also reported consistency between task success and satisfaction metrics between standard and automated summative usability testing but noted that having a usability engineer observe the sessions led to the discovery of a more comprehensive set of issues. Sauro (2009) found generally similar quantitative outcomes for completion rates and SUS scores from comparable lab-based and unmoderated usability testing, but noted the importance of removing "impossible" data collected with unmoderated testing (e.g., task times that were impossibly fast or unrealistically slow). In an aggregated meta-analysis of six published comparisons of lab-based and unmoderated remote usability studies, Sauro (2018d) reported very similar outcomes for completion rates and

ratings of ease-of-use, but more discrepancy in task completion times, with participants taking longer to finish in unmoderated studies.

In a study focused on problem discovery, Andreasen et al. (2007) reported similar outcomes between laboratory-based and synchronous remote usability testing but found fewer problems with asynchronous testing. In contrast, Bosenick et al. (2007) reported the discovery of more usability issues with remote asynchronous testing. Hertzum et al. (2015) categorized the relevance of verbalizations to the identification of usability issues in a comparison of relaxed moderated TA with unmoderated TA, reporting about equal percentages of medium and high relevance for the two experimental conditions and no important interactions of method with topic or valence (positive/neutral/negative). Hopefully, further research will reveal the reasons for these discrepant outcomes when comparing asynchronous usability testing to more standard laboratory-based testing, especially with regard to rates of problem discovery for the different methods.

3.5 Usability Laboratories

A typical usability laboratory test suite is a set of soundproofed rooms with a participant area and observer area separated by a one-way glass and with video cameras and microphones to capture the user experience (Marshall et al., 1990; Nielsen, 1997; Sauro, 2018e), possibly with an executive viewing area behind the primary observers' area. The advantages of this type of usability facility are quick setup, a place where designers can see people interacting with their products, videos to provide a historical record and backup for observers, and a professional appearance that raises awareness of usability and reassures customers about commitment to usability. Organizations rated as mature in UX are almost twice as likely to have a dedicated usability space as those rated as nonmature, and mature UX organizations are about three times more likely to use a one-way mirror in their usability labs (Sauro, 2018e).

In a survey of usability laboratories conducted in the mid-1990s, Nielsen (1994) reported a median floor space of 63 m^2 (678 ft^2) for the observer room and 13 m^2 (144 ft^2) for test rooms. Sauro (2018e) reported using between 11 and 15 m^2 (120 and 160 ft^2) for observer rooms and about 16 m^2 (168 ft^2) for test rooms, and that other considerations for usability labs include comfortable furniture, acoustic insulation, one-way mirror, remote-controlled video, microphones in the test rooms, push-to-talk mic in observation room to communicate with participants, direct high-speed Internet connections between observer and test rooms, computers and storage for captured audio/video, and a place for participants to wait. This type of laboratory is especially important if practitioners plan to conduct formal, summative usability tests.

If the practitioner focus is on formative, diagnostic problem discovery, this type of laboratory is not essential (although it is still convenient). "It is possible to convert a regular office temporarily into a usability laboratory, and it is possible to perform usability testing with no more equipment than a notepad" (Nielsen, 1997, p. 1561). Making an even stronger statement against the perceived requirement for formal laboratories, Landauer (1997, p. 204) stated:

> Many usability practitioners have demanded greater resources and more elaborate procedures than are strictly needed for effective guidance—such as expensive usability labs rather than natural settings for test and observations, time consuming videotaping and analysis where observation and note-taking would serve as well, and large groups of participants to achieve statistical significance when qualitative naturalistic observation of task goals and situations, or of disastrous interface or functionality flaws, would be more to the point.

In addition to remote usability testing (discussed above), another alternative to a formal, fixed-location usability laboratory is a mobile laboratory (Seffah & Habieb-Mammar, 2009). Advantages of mobile usability laboratories include portability to a participant's workplace and reduced cost relative to fixed laboratories. Because the mobile usability laboratory moves to the participant, disadvantages include the need to reduce the size of the usability testing team and complications in allowing nonteam observers to view the test.

3.6 Test Roles

There are several ways to categorize the roles that testers need to play in the preparation and execution of a usability test (Rubin, 1994; Dumas & Redish, 1999). Most test teams will not have a person assigned to each role, and most tests (especially informal problem discovery tests) do not require every role. The actual distribution of skills across a team might vary from these roles, but the standard roles help to organize the skills needed for effective usability testing.

3.6.1 Test Administrator

The test administrator is the usability test team leader. He or she designs the usability study, including the specification of the initial conditions for a test session and the codes to use for data logging. The test administrator's duties include conducting reviews with the rest of the test team, leading in the analysis of data, and putting together the final presentation or report. People in this role should have a solid understanding of the basics of usability engineering, the ability

to tolerate ambiguity, flexibility (knowing when to deviate from the plan), and good communication skills.

3.6.2 Briefer

The briefer is the person who interacts with the participants (briefing them at the start of the test, communicating with them as required during the test, and debriefing them at the end of the test sessions). On many teams, the same person takes the roles of administrator and briefer. In a TA study, the briefer has the responsibility to keep the participant talking. The briefer needs to have sufficient familiarity with the product to be able to decide what to tell participants when they ask questions. People in this role need to be comfortable interacting with people and need to be able to restrict their interactions to those that are consistent with the purposes of the test without any negative treatment of the participants.

3.6.3 Data Recorder

The video record is useful as a data backup when things start happening quickly during the test and as a source for video examples when documenting usability problems. The primary data source for a usability study, however, is the notes that the data recorder takes during a test session. There just is not time to take notes from a more leisurely examination of the video record. Also, the camera does not necessarily catch the important action at every moment of a usability study.

For informal studies, the equipment used to record data might be nothing more than a notepad and pencil. Alternatively, the data recorder might use data-logging software to take coded notes (often time stamped, possibly synchronized with the video). Before the test begins, the data recorder needs to prepare the data-logging software with the category codes defined by the test administrator. Taking notes with data-logging software is a very demanding skill, so the test administrator does not usually assign additional tasks to the person taking this role.

3.6.4 Product Expert

The product expert maintains the product and offers technical guidance during the test. The product expert must have sufficient knowledge of the product to recover quickly from product failures and to help the other team members understand the system’s actions during the test.

3.6.5 Statistician

A statistician has expertise in measurement and the statistical analysis of data. Practitioners with an educational background in experimental psychology typically have enough expertise to take the role of statistician for a usability test team. Informal tests rarely require the services of a statistician, but the team needs a statistician to extract the maximum amount of information from the data gathered during a formal test (especially if the purpose of the formal test was to compare two products using a battery of measurements).

3.7 Planning the Test

One of the first activities that a test administrator must undertake is to develop a test plan. To do this, the administrator must understand the purpose of the product, the parts of the product that are ready for test, the types of people who will use the product, what they are likely to use the product for, and in what settings.

3.7.1 Purpose of the Test

At the highest level, is the primary purpose of the test to identify usability problems or to gather usability measurements? The answer to this question provides guidance as to whether the most appropriate test is formal or informal, TA or silent, problem discovery or quantitative measurement. After addressing this question, the next task is to define any more specific test objectives. For example, an objective for an interactive voice response (IVR) system might be to assess whether participants can accomplish key tasks without encountering significant problems. If data is available from a previous study of a similar IVR, an alternative objective might be to determine whether participants can complete key tasks reliably faster with the new IVR than they did with the previous IVR. Most usability tests will include several objectives.

If a key objective of the test is to compare two products, an important decision is whether the test will be within subjects or between subjects. In a within-subjects test, every participant works with both products, with half of the participants using one product first and the other half using the other product first (a technique known as *counterbalancing*). In a between-subjects study, the test groups are completely independent. In general, a within-subjects test leads to more precise measurement of product differences (requiring a smaller number of participants for equal precision, due primarily to the reduction in variability that occurs when each participant acts as their own control) and the opportunity to get direct subjective product comparisons from participants. Tohidi et al. (2006) reported that participants exposed to alternative design solutions (within subjects) were more likely to provide informative criticism of the designs than participants

who worked with only one of the designs (between subjects). For a within-subjects test to be feasible, both products must be available and set up for use in the lab at the same time, and the amount of time needed to complete tasks with both products must not be excessive. If a within-subjects test is not possible, a between-subjects test is a perfectly valid alternative. Note that the statistical analyses appropriate for these two types of tests are different (for details, see Sauro & Lewis, 2016).

3.7.2 Participants

Representativeness

To determine who will participate in the test, the administrator needs to obtain or develop a user profile. A user profile is sometimes available from the marketing group, the product's functional specification, or other product planning documentation. It is important to keep in mind that the focus of a usability test is the end user of a product, not the expected product purchaser (unless the product will be purchased by end users). The most important participant characteristic is that the participant is representative of the population of end users to whom the administrator wants to generalize the results of the test. Practitioners can obtain participants from employment agencies, internal sources if the participants meet the requirements of the user profile (but avoiding internal test groups), market research firms, existing customers, colleges, online classified ads, and user groups.

To define representativeness, it is important to specify the characteristics that members of the target population share but are not characteristic of nonmembers. The administrator must do this for the target population at large and any defined subgroups. Within group definition constraints, administrators should seek heterogeneity in the final sample to maximize the generalizability of the results (Chapanis, 1988; Landauer, 1997) and to maximize the likelihood of problem discovery. It is true that performance measurements made with a homogeneous sample will almost always have greater precision than measurements made with a heterogeneous sample, but the cost of that increased precision is limited generalizability. This raises the issue of how to define homogeneity and heterogeneity of participants. After all, at the highest level of categorization, we are all humans, with similar general capabilities and limitations (physical and cognitive). At the other end of the spectrum, we are all individuals—no two alike.

One of the most important defining characteristics for a group in a usability test is specific relevant experience, both with the product and in the domain of interest (work experience, general product experience, specific product experience, experience with the product under test, and experience with similar products). One common categorization scheme is to consider people with less than

three months' experience as novices, with more than a year of experience as expert, and those in between as intermediate (Dumas & Redish, 1999). By definition, experts differ from novices with regard to effectiveness and efficiency of performance, at least partly due to changes in motor skills and usage strategies over time (Norman, 1983; Rasmussen, 1986). UX researchers have also consistently found that users with more product experience (more years of experience or more frequent use) report greater levels of perceived usability (Kortum & Bangor, 2013; Kortum & Johnson, 2013; Kortum & Sorber, 2015; Lah & Lewis, 2016; McLellan et al., 2012). UX studies of novices usually turn up more usability problems than studies of experts (Sauro, 2018c).

Fisher (1991) emphasized the importance of discriminating between computer experience (which he placed on a novice–experienced dimension) and domain expertise (which he placed on a naïve–expert dimension). LaLomia and Sidowski (1990) reviewed the scales and questionnaires developed to assess computer satisfaction, literacy, and aptitudes. None of the instruments they surveyed specifically addressed measurement of computer experience. Arning and Ziefle (2008) published an 18-item computer expertise questionnaire for older adults which assesses both theoretical computer knowledge and practical computer knowledge.

Other individual differences that practitioners routinely track and attempt to vary are education level, age, and gender, partly because they are relatively easy to define. It isn't clear, however, whether they usually have much effect on measures of usability and UX. For example, Billestrup et al. (2016), working with skilled Internet users, found that gender, age, and background (job function and education) did not have much effect on the usability problems participants experience. Most retrospective user experience studies using the System Usability Scale have found no significant effect of gender or age (Lewis, 2018c). Sonderegger et al. (2015) found no difference between older and younger users with regard to percentage of successful task completions, but also found that younger users tended to complete tasks more quickly.

When acquiring participants, how can practitioners define the similarity between the participants they can acquire and the target population? An initial step is to develop a taxonomy of the variables that affect human performance (where performance should include the behaviors of indicating preference and other choice behaviors). Gawron et al. (1989) produced a human performance taxonomy during the development of a human performance expert system. They reviewed existing taxonomies and filled in some missing pieces. They structured the taxonomy as having three top levels: environment, subject (person), and task. The resulting taxonomy took up 12 pages in their paper and covered many areas that would normally not concern a usability practitioner working in the field of computer system usability (e.g., ambient vapor pressure, gravity, acceleration).

Some of the key human variables in the Gawron et al. (1989) taxonomy that could affect human performance with computer systems are:

- Physical characteristics
  - Age
  - Agility
  - Handedness
  - Voice
  - Fatigue
  - Gender
  - Body and body part size
- Mental state
  - Attention span
  - Use of drugs (both prescription and illicit)
  - Long-term memory (includes previous experience)
  - Short-term memory
  - Personality traits
  - Work schedule
- Senses
  - Auditory acuity
  - Tone perception
  - Tactual
  - Visual accommodation
  - Visual acuity
  - Color perception

These variables can guide practitioners as they attempt to describe how participants and target populations are similar or different. The Gawron et al. (1989) taxonomy, however, does not provide much detail with regard to some individual differences that other researchers have hypothesized can affect human performance or preference with respect to the use of computer systems: personality traits and computer-specific experience.

Aykin and Aykin (1991) performed a comprehensive review of the published studies to that date that involved individual differences in human–computer interaction (HCI). Table 1 lists the individual differences that they found in published HCI studies, the method used to measure the individual difference, and whether there was any indication from the literature that manipulation of that individual difference led to a crossed interaction.

Table 1 Results of Aykin and Aykin (1991) Review of Individual Differences in HCI

Individual difference	Measurement method	Crossed interactions
Level of experience	Various methods	No
Jungian personality types	Myers–Briggs type of indicator	No
Field dependence/ independence	Embedded figures test	Yes; field-dependent participants preferred organized sequential item number search mode, but field-independent subjects preferred the less organized keyword search mode (Fowler et al., 1985)
Locus of control	Levenson test	No
Imagery	Individual differences questionnaire	No
Spatial ability	VZ-2	No
Type A/type B personality	Jenkins activity survey	No
Ambiguity tolerance	Ambiguity tolerance scale	No
Gender	Unspecified	No
Age	Unspecified	No
Other (reading speed and comprehension, intelligence, mathematical ability)	Unspecified	No

In statistical terminology, an interaction occurs whenever an experimental treatment has a different magnitude of effect depending on the level of a different, independent experimental treatment. A crossed interaction occurs when the magnitudes have different signs, indicating reversed directions of effects. As an example of an uncrossed interaction, consider the effect of turning off the lights on the typing throughput of blind and sighted typists. The performance of the sighted typists would probably be worse, but the presence or absence of light should not affect the performance of the blind typists. As an extreme example of a crossed interaction, consider the effect of language on task completion for people fluent only in French or English. When reading French text, French speakers would outperform English speakers, and vice versa.

For any of these individual differences, the lack of evidence for crossed interactions could be due to a paucity of research involving the individual difference or could reflect the probability that individual differences will not typically cause crossed interactions in HCI. In general, a change made to support a problem experienced by a person with a particular individual difference will either help other users or simply not affect their performance.

For example, John Black (personal communication, 1988) cited the difficulty that field-dependent users had working with one-line editors at the time (decades ago) when that was the typical user interface to a mainframe computer. Switching to full-screen editing resulted in a performance improvement for both field-dependent and field-independent users—an uncrossed interaction because both types of users improved, with the performance of field-dependent users becoming equal to (thus improving more than) that of field-independent users. Landauer (1997) cites another example of this, in which Greene et al. (1986) found that young people with high scores on logical reasoning tests could master database query languages such as SQL with little training, but older or less able people could hardly ever master these languages. They also determined that an alternative way of forming queries, selecting rows from a truth table, allowed almost everyone to make correct specification of queries, independent of their abilities. Because this redesign improved the performance of less able users without diminishing the performance of the more able, it was an uncrossed interaction. Palmquist and Kim (2000) found that field dependence affected the search performance of novices using a Web browser (with field-independent users searching more efficiently) but did not affect the performance of more experienced users.

Kortum and Oswald (2017) investigated the impact of personality on SUS ratings of perceived usability. People's scores on personality traits have been shown to be reliable and to predict important outcomes in work, school, and life domains. In this study, 268 participants used the SUS to retrospectively assess the perceived usability of 20 different products. Participants also completed a personality inventory which provides measurement of five broad personality traits: Extroversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience. There were significant correlations between the SUS and measures of Openness to Experience and Agreeableness. It is important for researchers and practitioners to be aware of potential effects of personality traits on the assessment of perceived usability, and to ensure that sample selection procedures, as appropriate, are unbiased with regard to the traits of Openness to Experience and Agreeableness. It seems unlikely that practitioners would routinely require participants to complete a personality inventory, but there may be value for researchers to include this step to replicate and extend the work of Kortum and Oswald (2017), furthering our understanding of these effects.

If there is a reason to suspect that an individual difference will lead to a crossed interaction as a function of interface design, it could make sense to invest the time (which can be considerable) to categorize users according to these dimensions. Another situation in which it could make sense to invest the time in categorization by individual difference would be if there were reasons to believe that a change in interface would greatly help one or more groups without adversely affecting other groups. (This is a strategy that one can employ when developing hypotheses about ways to improve user interfaces.) It always makes sense to keep track of user characteristics when categorization is easy (e.g., age or gender). Another potential use of these types of variables is as covariates (used to reduce estimates of variability) in advanced statistical analyses (Cliff, 1987).

Aykin and Aykin (1991) reported effects of users' levels of experience but did not report any crossed interactions related to this individual difference. They did report that interface differences tended to affect the performance of novices but had little effect on the performance of experts. It appears that behavioral differences related to user interfaces (Aykin & Aykin, 1991) and cognitive style (Palmquist & Kim, 2000) tend to fade with practice. Nonetheless, user experience has been one of the few individual differences to receive considerable attention in HCI research (Fisher, 1991; Mayer, 1997; Smith et al., 1999). According to Mayer (1997), relative to novices, experts have (1) better knowledge of syntax; (2) an integrated conceptual model of the system; (3) more categories for more types of routines; and (4) higher-level plans.

One user characteristic not addressed in any of the literature cited is one that becomes very important when designing products for international use: cultural characteristics. For example, in adapting an interface for use by members of another country, it is extremely important that all text be translated accurately. It is also important to be sensitive to the possibility that these types of individual differences might be more likely than others to result in crossed interactions.

For comparison studies, having multiple groups (e.g., males and females or experts and novices) allows the assessment of potential interactions that might otherwise go unnoticed. Ultimately, the decision for one or multiple groups must be based on expert judgment and a few guidelines. For example, practitioners should consider sampling from different groups if they have reason to believe:

- There are potential and important differences among groups on key measures (Dickens, 1987).
- There are potential interactions as a function of group (Aykin and Aykin, 1991).
- The variability of key measures differs as a function of the group.
- The cost of sampling differs significantly from group to group.

Gordon and Langmaid (1988) recommended the following approach to defining groups:

1. Write down all the important variables.
2. If necessary, prioritize the list.
3. Design an ideal sample.
4. Apply common sense to collapse cells.

For example, suppose that a practitioner starts with 24 cells, based on the factorial combination of six demographic locations, two levels of experience, and the two types of gender. Practitioners should ask themselves whether there is a high likelihood of learning anything new and important after completing the first few cells or whether additional testing would be wasteful. Can one learn just as much from having one or a few cells that are homogeneous within cells and heterogeneous between cells with respect to an important variable but are heterogeneous within cells with regard to other, less important variables? For example, a practitioner might plan: (1) to include equal numbers of males and females over and under 40 years of age in each cell; (2) to have separate cells for novice and experienced users; and (3) to drop intermediate users from the test. The resulting design requires testing only two cells (groups), but a design that did not combine genders and age groups in the cells would have required eight cells.

Sample Size

The final issue is the number of participants to include in the test. According to Dumas and Redish (1999), typical usability tests have 6–12 participants divided among 2–3 subgroups. Sauro (2010a) reported that the majority of industrial usability tests had total sample sizes in the range of 8–12 participants. For any given test, the required sample size depends on the number of subgroups, available resources (time/money), and purpose of the test (e.g., precise measurement or problem discovery). It also depends on whether a study is single shot (needing a larger sample size) or iterative (needing a smaller sample size per iteration, building up the total sample size over iterations) (Sauro & Lewis, 2016). Regarding sample sizes for problem discovery, Dumas (2003, p. 1098) wrote:

> This research does not mean that all of the *possible* problems with a product appear with 5 or 10 participants, but most of the problems that are going to show up with one sample of tasks and one group of participants will occur early.

Although these types of general guidelines have been helpful, it is possible to use more precise methods to estimate sample size requirements for problem discovery usability tests (Lewis, 1982, 1994, 2001, 2006; Turner, Lewis, & Nielsen, 2006). Estimating sample sizes for tests that have the primary purpose of discovering the problems in an interface depends on having an estimate of p, characterized as the average likelihood of problem occurrence or, alternatively, the problem discovery rate. As with comparative studies, this estimate can come from previous studies

using the same method and similar system under evaluation or can come from a pilot study.

For standard scenario-based usability studies, the literature contains large-sample examples that show *p* ranging from 0.08 to 0.46 (Hwang & Salvendy, 2007, 2009, 2010; Lewis, 1994; Nielsen & Landauer, 1993; Virzi, 1990, 1992). For heuristic evaluations, the reported value of *p* from large-sample studies ranges from 0.08 to 0.60 (Hwang & Salvendy, 2007, 2009, 2010; Nielsen & Molich, 1990). The well-known (and often misused and maligned) guideline that five participants are enough to discover 85% of problems available for discovery is true only when $p = 0.315$. As the reported ranges of *p* indicate, there will be many studies for which this guideline (or any similar guideline) will not apply, making it important for usability practitioners either to obtain estimates of *p* for their usability studies or to use the information in Table 2 to understand the relationship among *p*, sample size, and the cumulative likelihood of problem discovery.

The cells in Table 2 are the probability of having a problem with a specified probability of occurrence happen *at least once* during a usability study with the given sample size (using the formula $1 - (1—p)^n$ after solving for *n*—for details, see Sauro & Lewis, 2016). Practitioners who are uncomfortable with sample size estimation procedures that implicitly assume a fixed number of problems available for discovery (Hornbæk, 2010) or are concerned with unmodeled variability of an averaged estimate of *p* (Borsci, MacRedie, Barnett, Martin, Kuljis, & Young, 2013; Caulton, 2001; Schmettow, 2008, 2012; Woolrych & Cockton, 2001) can use Table 2 to plan their formative usability studies without those limitations. Sauro (2019b) analyzed problem discovery from seven industrial usability evaluations and verified that the magnitude of problem discovery with the first five participants matched the expected discovery rate based on the entire sample for different likelihoods of problem occurrence.

One of the first published studies of problem discovery as a function of sample size reported that severe problems were likely to occur with the first few participants (Virzi, 1992). However, there is nothing in $1 - (1—p)^n$ that would account for anything other than the probable frequency of occurrence as influencing early appearance of an event of interest in a user study (Lewis, 1994). Law and Hvannberg (2004) reported no significant correlation between problem severity and problem detection rate. Sauro (2014) studied the problem severity ratings of multiple evaluators across nine usability studies independently using their judgment, as opposed to data-driven assessments, and found that the average correlation across all nine studies was not significantly different from zero (only one study showed a significant positive correlation). Although a few studies have indicated a positive correlation between problem frequency and severity, the preponderance of the data indicates that there is no reliable relationship, so the best policy is for practitioners to assume no relationship when planning usability studies.

Table 2 Likelihood of Discovering Problems of Probability *p* at Least Once in a Study with Sample Size *n*

	Sample Size						
p	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
0.05	0.10	0.14	0.19	0.23	0.26	0.30	0.34
0.10	0.19	0.27	0.34	0.41	0.47	0.52	0.57
0.15	0.28	0.39	0.48	0.56	0.62	0.68	0.73
0.25	0.44	0.58	0.68	0.76	0.82	0.87	0.90
0.50	0.75	0.88	0.94	0.97	0.98	0.99	1.00
0.90	0.99	1.00	1.00	1.00	1.00	1.00	1.00
p	**9**	**10**	**11**	**12**	**13**	**14**	**15**
0.01	0.09	0.10	0.10	0.11	0.12	0.13	0.14
0.05	0.37	0.40	0.43	0.46	0.49	0.51	0.54
0.10	0.61	0.65	0.69	0.72	0.75	0.77	0.79
0.15	0.77	0.80	0.83	0.86	0.88	0.90	0.91
0.25	0.92	0.94	0.96	0.97	0.98	0.98	0.99
0.50	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00
p	**16**	**17**	**18**	**19**	**20**	**25**	**30**
0.01	0.15	0.16	0.17	0.17	0.18	0.22	0.26
0.05	0.56	0.58	0.60	0.62	0.64	0.72	0.79
0.10	0.81	0.83	0.85	0.86	0.88	0.93	0.96
0.15	0.93	0.94	0.95	0.95	0.96	0.98	0.99
0.25	0.99	0.99	0.99	1.00	1.00	1.00	1.00
0.50	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00

3.7.3 Test Task Scenarios

As with participants, the most important consideration for test tasks is that they are representative of the types of tasks that real users will perform with the product. For any product, there will be a core set of tasks that anyone using the product will perform. People who use barbecue grills, use them to cook. People who use desktop speech dictation products, use them to produce text. People who use their banking website, check their balances and transfer money. For usability tests, these are the most important tasks to test.

After defining these core tasks, the next step is to list any more peripheral tasks that the test should cover. If a barbecue grill has an external burner for heating pans, it might make sense to include a task that requires participants to

work with that burner. If in addition to the basic vocabulary in a speech dictation system the program allows users to enable additional special topic vocabularies such as cooking or sports, it might make sense to devise a task that requires participants to activate and use one of these topics. Practitioners should avoid frivolous or humorous tasks because what is humorous to one person might be offensive or annoying to another.

From the list of test tasks, create scenarios of use (with specific goals) that require participants to perform the identified tasks. Critical tasks can appear in more than one scenario. For repeated tasks, vary the task details to increase the generalizability of the results. When testing relatively complex systems, some scenarios should stay within specific parts of the system (e.g., typing and formatting a document) and others should require the use of different parts of the system (e.g., creating a figure using a spreadsheet program, adding it to the document, attaching the document to a note, and sending it to a specified recipient).

The complete specification of a scenario should include several items. It is important to document (but not to share with the participant) the required initial conditions so it will be easy to determine before a test session starts if the system is ready and the required ending conditions that define successful task completion (Howard, 2008; Howard & Howard, 2009). The written description of the scenario (presented to the participant) should state what the participant is trying to achieve and why (the motivation), keeping the description of the scenario as short as possible to keep the test session moving quickly. The scenario should end with an instruction for the action the participant should take upon finishing the task (to make it easier to measure task completion times). The descriptions of the scenario's tasks should not typically provide step-by-step instructions on how to complete the task but should include details (e.g., actual names and data) rather than general statements. For tasks in which users work with highly personalized data (email, calendar, financial), scenarios constructed with a participant's own real data can increase the validity of the study (Genov, Keavney, & Zazelenchuk, 2009).

The order in which participants complete scenarios should reflect the way in which users would typically work and with the importance of the scenario, with important scenarios done first unless there are other less important scenarios that produce outputs that the important scenario requires as an initial condition. Within those constraints, try to vary their order of presentation. Not all participants need to receive the same scenarios, especially if there are different groups under study. The tasks performed by administrators of a Web system that manages subscriptions will be different from the tasks performed by users who are requesting subscriptions.

Here are some examples of scenarios:

- Frank Smith's business telephone number has changed to (896) 555-1234. Please change the appropriate address book entry so you have this new phone number available when you need it. When you have finished, please say "I'm done."
- You've just found out that you need to cancel a car reservation that you made for next Wednesday. Please call the system that you used to make the reservation (1-888-555-1234) and cancel it. When you have finished, please hang up the phone and say, "I'm done."

Bailey et al. (2009) have described stopping a task after the first step as a way of assessing a large number of tasks in a relatively short period of time. Over a number of website studies, they found that if the first click of a task was correct, the likelihood of final task success was 0.87, whereas if the first click was incorrect, the likelihood of final success was 0.46. In a replication first-click study, Sauro (2013) found a smaller difference in final task success for tasks that started with a searching phase, but a much larger difference (80% to 14%) for navigation-only tasks. The more tasks covered in a usability test, the greater the likelihood of discovery of usability problems (Lindgaard & Chattratichart, 2007).

3.7.4 Procedure

The test plan should include a description of the procedures to follow when conducting a test session. Most test sessions include an introduction, task performance, posttask activities, and debriefing.

A common structure for the introduction is for the briefer (review Section 3.6.2) to start with the purpose of the test, emphasizing that its goal is to improve the product, not to test the participant. Participation is voluntary, and the participant can stop at any time without penalty. The briefer should inform the participant that all test results will be confidential. The participant should be aware of any planned audio or video recording. Finally, the briefer should provide any special instructions (e.g., TA instructions) and answer any other questions that the participant might have.

The participant should then complete any preliminary questionnaires and forms, such as a background questionnaire, an informed consent form (including consent for any recording, if applicable), and, if necessary, a confidential disclosure form. If the participant will be using a workstation, the briefer should help the participant make any necessary adjustments (unless, of course, the purpose of the test is to evaluate workstation adjustability). Finally, the participant should complete any prerequisite training. This can be especially important if the goal of the study is to investigate usability after some period of use rather than immediate usability.

The procedure section should indicate the order in which participants will complete task scenarios. For each participant, start with the first task scenario assigned and complete additional scenarios until the participant finishes (or runs out of time). The procedure section should specify when and how to interact with participants, according to the type of study. This section should also indicate when it is permissible to assist participants if they encounter difficulties in task performance.

Normally, practitioners should avoid offering assistance unless the participant is visibly distressed. When participants initially request help at a given step in a task, refer them to documentation or other supporting materials if available. If that doesn't help, provide the minimal assistance required to keep the participant moving forward in the task, note the assistance, and score the task as failed. When participants ask questions, try to avoid direct answers, instead turning their attention back to the task and encouraging them to take whatever action seems right at that time. When asking questions of participants, it is important to avoid biasing the participant's response. Try to avoid the use of loaded adjectives and adverbs in post-task interviews (Dumas & Redish, 1999). Instead of asking if a task was easy, ask the participant to describe what it was like performing the task. Assess perceived usability using a standard method (such as the SEQ; see Section 3.9.7 for details) at the end of each scenario.

After participants have finished the assigned scenarios, it is common to have them complete a final questionnaire, usually a standard questionnaire and any additional items required to cover other test- or product-specific issues. For standardized questionnaires, ISO lists the SUMI (Software Usability Measurement Inventory) (Kirakowski, 1996; Kirakowski & Corbett, 1993) and PSSUQ (Post-Study System Usability Questionnaire) (Lewis, 1995, 2002). In addition to the SUMI and PSSUQ, ANSI lists the QUIS (Questionnaire for User Interaction Satisfaction) (Chin, Diehl, & Norman, 1988) and SUS (System Usability Scale) (Brooke, 1996) as widely used questionnaires. See Section 3.9 in this chapter for descriptions of these and more recently developed questionnaires.

After completing the final questionnaire, the briefer should debrief the participant. Toward the end of debriefing, the briefer should tell the participant that the test session has turned up several opportunities for product improvement (this is almost always true) and thank the participant for their contribution to product improvement. Finally, the briefer should discuss any questions that the participant has about the test session and then take care of any remaining activities. If any deception has been employed in the test (which is rare but can happen legitimately when conducting certain types of simulations), the briefer has an ethical obligation to inform the participant.

3.7.5 Pilot Testing

Practitioners should always plan for a pilot test before running a usability test. A usability test is a designed artifact and like any other designed artifact needs at least some usability testing to find problems in the test procedures and materials. A common strategy is to have an initial walkthrough with a member of the usability test team or some other convenient participant. After making the appropriate adjustments, the next pilot participant should be a more representative participant. If there are no changes made to the design of the usability test after running this participant, the second pilot participant can become the first real participant (but this is rare). Pilot testing should continue until the test procedures and materials have become stable.

3.7.6 Number of Iterations

It is better to run one usability test than not to run any at all. On the other hand, "usability testing is most powerful and most effective when implemented as part of an iterative product development process" (Rubin, 1994, p. 30). Ideally, usability testing should begin early and occur repeatedly throughout the development cycle. When development cycles are short, it is a common practice to run, at a minimum, exploratory usability tests on prototypes at the beginning of a project, to run a usability test on an early version of the product during the later part of functional testing, and then to run another during system testing. Once the final version of the product is available, some organizations run an additional usability test focused on the measurement of usability performance benchmarks. At this stage of development, it is too late to apply information about any problems discovered during the usability test to the soon-to-be-released version of the product, but the information can be useful as early input to a follow-on product if the organization plans to develop another version of the product.

3.7.7 Ethical Treatment of Test Participants

Usability testing always involves human participants, so usability practitioners must be aware of professional practices in the ethical treatment of test participants. Practitioners with professional education in experimental psychology are usually familiar with the guidelines of the American Psychology Association (APA; see http://www.apa.org/ethics/), and those with training in human factors engineering are usually familiar with the guidelines of the Human Factors and Ergonomics Society (HFES) (see https://www.hfes.org/about-hfes/code-of-ethics). It is particularly important (Dumas, 2003) to be aware of the concepts of informed consent (participants are aware of what will happen during the test, agree to participate, and can leave the test at any time without penalty) and minimal risk

(participating in the test does not place participants at any greater risk of harm or discomfort than situations normally encountered in daily life). Most usability tests are consistent with guidelines for informed consent and minimal risk. Only the test administrator should be able to match a participant's name and data, and the names of test participants should be confidential. Anyone interacting with a participant in a usability test has a responsibility to treat the participant with respect.

Usability practitioners rarely use deception in usability tests. One technique in which there is potential use of deception is the WoZ method (originally, the OZ Paradigm) (Kelley, 1985, 2018). In a test using the WoZ method, a human (the Wizard) plays the part of the system, remotely controlling what the participant sees happen in response to the participant's manipulations. This method is particularly effective in early tests of speech recognition IVR systems because all the Wizard needs is a script and a phone (Sadowski, 2001). Often, there is no compelling reason to deceive participants, so they know that the system they are working with is remotely controlled by another person for the purpose of early evaluation. If there is a compelling need for deception (e.g., to manage the participant's expectations and encourage natural behaviors), this deception must be revealed to the participant during debriefing.

3.8 Reporting Results

There are two broad classes of usability test results, problem reports and quantitative measurements. It is possible for a test report to contain one type exclusively (e.g., the ANSI Common Industry Format has no provision for reporting problems, which led the National Institute of Standards and Technology to investigate a similar standard for formative test reports; see Theofanos & Quesenbery, 2005), but most usability test reports will contain both types of results. Høegh et al. (2006) reported that usability reports can have a strong impact on developers' understanding of specific usability problems, especially if the developers have also observed usability test sessions. Of particular interest to the developers was the list of specific usability problems and redesign proposals, consistent with the results of Capra (2007) and Nørgaard and Hornbæk (2009).

3.8.1 Describing Usability Problems

According to Marshall et al.:

> We broadly define a usability defect as: Anything in the product that prevents a target user from achieving a target task with reasonable effort and within a reasonable time. … Finding usability problems is relatively easy. However, it is

> much harder to agree on their importance, their causes and the changes that should be made to eliminate them (the fixes).
>
> *(1990, p. 245)*

The best way to describe usability problems depends on the purpose of the descriptions. For usability practitioners, the goal should be to describe problems in such a way that the description leads logically to one or more potential interventions (recommendations) that will help designers and developers improve the system under evaluation (Høegh et al., 2006; Hornbæk, 2010). Ideally, the problem description should also include some indication of the importance of fixing the problem, most often referred to as problem severity. For more scientific investigations, there can be value in higher levels of problem description (Keenan, Hartson, Kafura, & Schulman, 1999), but developers rarely care about these levels of description. They just want to know what they need to do to make things better while also managing the cost, both monetary and time, of interventions (Gray & Salzman, 1998).

The problem description scheme of Lewis and Norman (1986) has both scientific and practical merit because their problem description categories indicate, at least roughly, an appropriate intervention. They stated (p. 413) that "although we do not believe it possible to design systems in which people do not make errors, we do believe that much can be done to minimize the incidence of error, to maximize the discovery of the error, and to make it easier to recover from the error." They separated errors into mistakes (errors due to incorrect intention) and *slips* (errors due to appropriate intention but incorrect action), further breaking slips down into *mode errors* (which indicate a need for better feedback or elimination of the mode), *capture errors* (which indicate a need for better feedback), and *description errors* (which indicate a need for better design consistency). In one study using this type of problem categorization, Prümper et al. (1992) found that expertise did not affect the raw number of errors made by participants in their study, but experts handled errors much more quickly than novices. The types of errors that experts made were different from those made by novices, with experts' errors occurring primarily at the level of slips rather than mistakes.

Using an approach similar to that of Lewis and Norman (1986), Rasmussen (1986) described three levels of errors: (1) skill-based; (2) rule-based; and (3) knowledge-based. Other classification schemes include Structured Usability Problem Extraction, or SUPEX (Cockton & Lavery, 1999), the User Action Framework, or UAF (Andre, Belz, McCreary, & Hartson, 2000), and the Classification of Usability Problems (CUP) scheme (Vilbergsdóttir, Hvannberg, & Law, 2006). The UAF requires a series of decisions, starting with an interaction

cycle (planning, physical actions, assessment) based on the work of Norman (1986). Most classifications require four or five decisions, with interrater reliability, as measured with kappa (κ). highest at the first step ($\kappa = 0.978$) but remaining high through the fourth and fifth steps ($\kappa > 0.7$). Yusop et al. (2017) conducted a literature review of 57 studies (37 usability studies, 20 software engineering studies) and reported that the usability defect processes in those studies had a number of limitations, including mixed data, inconsistency of terms, and insufficient information for classification.

Whether any of these classification schemes will see widespread use by usability practitioners is still unknown. For example, the CUP scheme requires some training for inexperienced evaluators to effectively use the scheme, even though a simplified version may be useful for developers and usability practitioners (Vilbergsdóttir et al., 2006). There is considerable pressure on practitioners to produce results and recommendations as quickly as possible. Even if these classification schemes see little use by practitioners, effective problem classification is a very important problem to solve as usability researchers strive to compare and improve usability testing methods.

3.8.2 Crafting Design Recommendations from Problem Descriptions

The development of recommendations from problem descriptions is a craft rather than a rote procedure. A well-written problem description will often strongly imply an intervention, but it is also often the case that there might be several ways to attack a problem. It can be helpful for practitioners to discuss problems and potential interventions with the other members of their team and to get input from other stakeholders as necessary (especially, the developers of the product). This is especially important if the practitioner has observed problems but is uncertain as to the appropriate level of description of the problem.

For example, suppose that you have written a problem description about a missing Help button in a software application. This could be a problem with the overall design of the software or might be a problem isolated to one screen. You might be able to determine this by inspecting other screens in the software, but it could be faster to check with one of the developers.

The first recommendations to consider should be for interventions that will have the widest impact on the product. “Global changes affect everything and need to be considered first” (Rubin, 1994, p. 285). After addressing global problems, continue working through the problem list until there is at least one recommendation for each problem. For each problem, start with interventions that would eliminate the problem, then follow, if necessary, with other less drastic (less expensive, more likely to be implemented) interventions that would reduce the

severity of the remaining usability problem. When different interventions involve different tradeoffs, it is important to communicate this clearly in the recommendations. This approach can lead to two tiers of recommendations: those that will happen for the version of the product currently under development (short-term) and those that will happen for a future version of the product (long-term).

Molich et al. (2007) used results from CUE-4 to develop guidelines for making usability recommendations useful and usable. By their assessment, only 14 of 84 studied comments (17%) were both useful and usable. To address the weaknesses observed in the recommendations, they concluded:

- Communicate clearly at the conceptual level.
- Ensure that recommendations improve overall usability.
- Be aware of business or technical constraints.
- Solve the whole problem, not just a special case.

Nørgaard and Hornbæk (2009) conducted an exploratory study in which three developers assessed 40 usability findings presented using five feedback formats. The developers rated redesign proposals, multimedia presentations, and screen dumps as useful inputs, problem lists second, and scenarios as least helpful. "Problem lists seem best suited for communicating simple and uncontroversial usability problems for which no contextual information is needed" (p. 64). The preferred feedback formats provided strong contextual information. These results suggest that problem lists can be useful, but it is important to provide sufficient contextual information, if not possible through verbal description, then through associated redesign proposals, screen dumps, and multimedia presentations.

3.8.3 Prioritizing Problems

Because usability tests can reveal more problems than there are resources to address, it is important to have some means for prioritizing problems, keeping in mind that design process considerations (stage of development and cost-effectiveness) can also influence the specific usability changes made to a product (Hertzum, 2006). There are two approaches to prioritization that have appeared in the usability testing literature: (1) judgment-driven (Virzi, 1992); and (2) data-driven (Dumas & Redish, 1999; Lewis, Henry, & Mack, 1990; Rubin, 1994). The bases for judgment-driven prioritizations are the ratings of stakeholders in the project (such as usability practitioners and developers). The bases for data-driven prioritizations are the data associated with the problems, such as frequency, impact, ease of correction, and likelihood of usage of the portion of the product that was in use when the problem occurred. Of these, the most common measurements are frequency and impact (sometimes referred to as severity, although, strictly speaking, severity should include the effect of all of the types of data considered for prioritization). In a study of the two approaches to

prioritization, Hassenzahl (2000) found a lack of correspondence between data-driven and judgment-driven severity estimates. This suggests that the preferred approach should be data-driven.

The usual method for measuring the frequency of occurrence of a problem is to divide the number of occurrences within participants by the number of participants. A common method (Dumas & Redish, 1999; Rubin, 1994) for assessing the impact of a problem is to assign impact scores according to whether the problem (1) prevents task completion; (2) causes a significant delay or frustration; (3) has a relatively minor effect on task performance; or (4) is a suggestion. This is similar to the scheme of Lewis et al. (1990), in which the impact levels were:

1. scenario failure or irretrievable data loss (e.g., the participant required assistance to get past the problem or it caused the participant to believe the scenario to be properly completed when it was not);
2. considerable recovery effort (recovery took more than 1 min. or the participant repeatedly experienced the problem within a scenario);
3. minor recovery effort (the problem occurred only once within a scenario with recovery time at or under 1 min.);
4. inefficiency (a problem not meeting any of the other criteria).

When considering multiple types of data in a prioritization process, it is necessary to combine the data in some way. A graphical approach is to create a problem grid with frequency on one axis and impact on the other. High-frequency, high-impact problems would receive treatment before low-frequency, low-impact problems. The relative treatment of high-frequency, low-impact problems and low-frequency, high-impact problems depends on practitioner judgment.

An alternative approach is to combine the data arithmetically. Rubin (1994) described a procedure for combining four levels of impact (using the criteria described above with 4 assigned to the most serious level) with four levels of frequency (4: frequency ≥ 90%; 3: 51–89%; 2: 11–50%; 1: ≤ 10%) by adding the scores. For example, if a problem had an observed frequency of occurrence of 80% and had a minor effect on performance, its priority would be 5 (a frequency rating of 3 plus an impact rating of 2). With this approach, priority scores can range from a low of 2 to a high of 8. If information is available about the likelihood that a user would work with the part of the product that enables the problem, this information would be used to adjust the frequency rating. Continuing the example, if the expectation is that only 10% of users would encounter the problem, the priority would be 3 (a frequency rating of 1 for the 10% × 80%, or an 8% likelihood of occurrence plus an impact rating of 2).

A similar strategy is to multiply the observed percentage frequency of occurrence by the impact score. The range of priorities depends on the values assigned to each impact level. Assigning 10 to the most serious impact level leads to a maximum priority (severity) score of 1000 (which can optionally be divided by 10 to create a scale that ranges from 1 to 100). Appropriate values for the remaining three impact categories depend on practitioner judgment, but a reasonable set is 5, 3, and 1. Using those values, the problem with an observed frequency of occurrence of 80% and a minor effect on performance would have a priority of 24 (80 × 3/10). It is possible to extend this method to account for the likelihood of use using the same procedure as that described by Rubin (1994), which in the example resulted in modifying the frequency measurement from 80% to 8%. Another way to extend the method is to categorize the likelihood of use with a set of categories such as very high likelihood (assigned a score of 10), high likelihood (assigned a score of 5), moderate likelihood (assigned a score of 3), and low likelihood (assigned a score of 1) and multiply all three scores to get the final priority (severity) score (then optionally divide by 100 to create a scale that ranges from 1 to 100). Continuing the previous example with the assumption that the task in which the problem occurred has a high likelihood of occurrence, the problem's priority would be 12 (5 × 240/100). In most cases, applying the different data-driven prioritization schemes to the same set of problems should result in a very similar prioritization, but there has been no research published on this topic.

3.8.4 Working with Quantitative Measurements

The most common use of quantitative measurements is to characterize performance and preference variables by computing means, standard deviations, and ideally confidence intervals (Sauro & Lewis, 2016). Practitioners use these results to compare observed to target measurements when targets are available. When targets are not available, the results can still be informative, for example, for use as future target measurements or as relatively gross diagnostic indicators.

The failure to meet targets is an obvious diagnostic cue. A less obvious cue is an unusually large standard deviation. Landauer (1997) describes a case in which the times to record an order were highly variable. The cause for the excessive variability was that a required phone number was sometimes, but not always, available, which turned out to be an easy problem to fix. Because the means and standard deviations of time scores tend to correlate, one way to detect an unusually large variance is to compute the coefficient of variation by dividing the standard deviation by the mean or the normalized performance ratio by dividing the mean by the standard deviation (Moffat, 1990). Large coefficients of variation (or, correspondingly, small normalized performance ratios) are potentially indicative of the presence of usability problems.

3.9 Standardized UX Questionnaires

As opposed to indirectly assessing UX with performance (task completion success rates and times) or biometric measures (galvanic skin response or heart rate), the most direct way to quantitatively measure perceived usability and UX is with standardized questionnaires. Standardized measures offer many advantages to usability and UX practitioners. Specifically, standardized measurements provide objectivity, replicability, quantification, economy, communication, and scientific generalization (Nunnally, 1978). Comparisons of the reliability of standardized versus ad hoc (home-grown) usability and UX questionnaires consistently favor the use of standardized instruments (Bargas-Avila & Hornbæk, 2011; Hornbæk, 2006; Hornbæk & Law, 2007; Sauro & Lewis, 2009; Tullis & Stetson, 2004). The first published standardized usability questionnaires appeared in the late 1980s (Chin et al., 1988; Kirakowski & Dillon, 1988). Questionnaires focused on the measurement of computer satisfaction preceded these questionnaires (e.g., the Gallagher Value of MIS Reports Scale and the Hatcher and Diebert Computer Acceptance Scale) (see LaLomia & Sidowski, 1990, for a review), but those questionnaires were not applicable to scenario-based usability tests.

The most widely used of the first generation of standardized usability questionnaires are the QUIS (Chin et al., 1988), the SUMI (Kirakowski, 1996; Kirakowski & Corbett, 1993), the PSSUQ (Lewis, 1992, 1995, 2002), and the SUS (Brooke, 1996, 2013) and of these, the most popular is the SUS (Lewis, 2018c). The most common application of these questionnaires is at the end of a test (after completing a series of test scenarios), although they can also be used for retrospective evaluation in surveys (Grier et al., 2013). The longer standardized questionnaires typically have completion times of less than 10 min. (Dumas, 2003).

Post-study questionnaires are important instruments in the usability practitioner's toolbox, but they assess experience at a relatively high level. This can be a strength when comparing competitors or different versions of a product, but is a weakness when seeking more detailed diagnoses of problem areas in a user interface. To address this weakness, many practitioners perform a quick assessment of perceived usability immediately after participants complete each task or scenario using a standardized approach such as the After-Scenario Questionnaire (Lewis, 1991b) or the Single Ease Question (Sauro & Dumas, 2009; Sauro & Lewis, 2016; Tedesco & Tullis, 2006). A quick measure of perceived usability at the task level is not the same as specific problem identification, but tasks that score poorly on these types of measures can draw attention to problems and help with their prioritization.

The primary measures of the quality of standardized questionnaire are reliability (consistency of measurement) and validity (measurement of the

intended attribute) (Nunnally, 1978). There are several ways to assess reliability, including test–retest and split-half reliability. The most common method for the assessment of reliability is coefficient α, a measurement of internal consistency. Coefficient α can range from 0 (no reliability) to 1 (perfect reliability). Measures that can affect a person's future, such as IQ tests or college entrance exams, should have a minimum reliability of 0.90 (preferably, reliability greater than 0.95). For other research or evaluation, measurement reliability in the range of 0.70–0.80 is acceptable (Landauer, 1997; Nunnally, 1978).

A questionnaire's validity is the extent to which it measures what it claims to measure. Researchers commonly use the Pearson correlation coefficient to assess criterion-related validity (the relationship between the measure of interest and a different concurrent or predictive measure). These correlations do not have to be large to provide evidence of validity. For example, personnel selection instruments with validities as low as 0.30 or 0.40 can be large enough to justify their use (Nunnally, 1978). Another approach to validity is content validity, typically assessed through the use of factor analysis (which also helps questionnaire developers discover or confirm clusters of related items that can form reasonable subscales).

Regarding the appropriate number of scale steps, more scale steps are better than fewer scale steps, but with rapidly diminishing returns (Lewis, 2019b; Lewis & Erdinç, 2017). The reliability of individual items tends to be a monotonically increasing function of the number of steps (Nunnally, 1978). As the number of scale steps increase from 2 to 20, the increase in reliability is very rapid at first but tends to level off at about 7. After 11 steps there is little gain in reliability from increasing the number. Despite the concerns of some UX researchers, there is no evidence that a smaller number of response options is easier for respondents to complete, and there is evidence that having only three response options can be problematic (Lewis, 2019b, Sauro, 2019a). The number of steps in an item is very important for metrics based on a single item but is less important when computing measurements over a number of items (as in the computation of an overall or subscale score).

3.9.1 QUIS

The Questionnaire for User Interaction Satisfaction (QUIS, Chin et al., 1988; Shneiderman, 1987) is a product of the Human–Computer Interaction Lab at the University of Maryland. Its use requires the purchase of a license. Chin et al. (1988) evaluated several early versions of the QUIS (Versions 3–5). They reported an overall reliability (coefficient α) of 0.94 but did not report any subscale reliability.

The QUIS is currently at Version 7. This version includes demographic questions, an overall measure of reaction to the software, and 11 specific interface factors. The QUIS is available in two lengths, short (26 items) and long (71 items). The items are 0–9-point scales anchored with opposing adjective phrases (such as "confusing" and "clear" for the item "messages which appear on screen"). Although the item content is primarily focused on ratings of system attributes, some items in the factor for overall reaction to the software capture emotional reactions (e.g., terrible vs. wonderful, frustrating vs. satisfying, dull vs. stimulating), showing early attention to UX measures.

3.9.2 SUMI

The SUMI (Kirakowski, 1996; Kirakowski & Corbett, 1993) is a questionnaire with six subscales: global, efficiency, affect, helpfulness, control, and learnability. Its 50 items are statements (such as "The instructions and prompts are helpful") to which participants indicate that they agree, are undecided, or disagree. The SUMI has undergone a significant amount of psychometric development and evaluation to arrive at its current form. The results of studies that included significant main effects of system, SUMI scales, and their interaction support its validity (McSweeney, 1992; Wiethoff, Arnold, & Houwing, 1992). By virtue of its inclusion of an Affect subscale, the SUMI is a first-generation standardized questionnaire that extended its reach beyond pragmatic usability factors to provide some assessment of the emotional aspects of UX.

The reported reliabilities of the six subscales (measured with coefficient α) are:

- Global: 0.92
- Efficiency: 0.81
- Affect: 0.85
- Helpfulness: 0.83
- Control: 0.71
- Learnability: 0.82

One of the greatest strengths of the SUMI is the database of results that is available for the construction of interpretive norms. This makes it possible for practitioners to compare their results with those of similar products and tasks, as long as there are similar products and tasks in the database; Cavallin et al. (2007) reported a significant effect of tasks on SUMI scores. Another strength is that the SUMI is available in different languages such as UK English, American English, Italian, Spanish, French, German, Dutch, Greek, and Swedish. Like the QUIS, practitioners planning to use SUMI must purchase a license for its use, which includes questionnaires and scoring software. For an additional fee, a trained psychometrician at the HFRG will score the results and produce a report.

3.9.3 SUPR-Q and SUPR-Qm

The Standardized User Experience Percentile Rank Questionnaire (SUPR-Q, Sauro, 2015b), now in its second version, is a UX rating scale designed to measure perceptions of usability, credibility/trust, appearance, and loyalty for websites. Note that three of the four SUPR-Q factors are based on constructs other than standard usability. Like the SUMI, commercial use of the SUPR-Q requires the purchase of a license (measuringu.com/product/suprq).

The SUPR-Q provides relative rankings expressed as percentages, so a SUPR-Q percentile score of 50 is average (roughly half the websites evaluated in the past with the SUPR-Q have received better scores and half received worse). In addition to this global comparison, the SUPR-Q has a normative database with data from over 200 websites and over 4000 users across multiple industries, updated quarterly. Thus, the questionnaire can be used to generate reliable scores in benchmarking websites, and the normed scores can be used to understand how well a website scores relative to others in the database.

The current version of the SUPR-Q has eight items (derived from an initial pool of 33 items, themselves drawn from the UX and market research literature), with seven 5-point items (1 = "Strongly disagree"; 5 = "Strongly agree") and one 11-point item Likelihood to Recommend. The reported reliabilities for the SUPR-Q subscales (Sauro, 2015b), were:

- Usability (Items 1, 2): 0.88
- Trust (Items 3, 4): 0.85
- Loyalty (Items 5, 8): 0.64
- Appearance (Items 6, 7): 0.78
- Global (All items): 0.86

All the SUPR-Q scale reliabilities exceeded 0.70 except for Loyalty, which was 0.64. The global SUPR-Q scores correlated significantly with concurrently collected SUS scores (r = 0.75), as did all four subscales (Usability: 0.73; Trust: 0.39; Loyalty: 0.61; Appearance: 0.64). The factor structure has been replicated across three studies with data collected both during usability tests and retrospectively in surveys. In a study of 40 websites (n = 2513), the global SUPR-Q and its subscales discriminated well between the poorest and highest quality websites, with about equal discriminating power as the SUS (Sauro, 2015b).

Using advanced psychometric methods, Sauro and Zarolia (2017) developed a variant of the SUPR-Q, the SUPR-Qm, for measurement of the mobile app user experience. The first step was to identify appropriate content, considering items associated with published constructs such as utility, usability, intended usage, and future usage, then through item analysis to winnow that number down to a set of items that described the quality of the mobile application user experience that applied to a broad range of app categories. Rasch analysis was

used to assess the psychometric properties of items collected from four independent surveys (n = 1,046) with ratings on 174 unique apps. Sixteen items were identified that fit the model well.

For the final version of the 16-item SUPR-Qm, reliability estimates were high (alpha = 0.94), as was convergent validity, with significant correlations with the SUPR-Q (0.71), UMUX-LITE (0.74, see Section 3.9.6), and likelihood-to-recommend (LTR) (0.74). Scores on the SUPR-Qm correlated with the number of app reviews in the Google Play Store and Apple's App Store (r = 0.38), establishing adequate predictive validity. The SUPR-Qm can be used to benchmark the user experience of mobile applications, an essential step in understanding what works and what needs improvement in mobile apps.

3.9.4 SUS

The System Usability Scale (SUS) is a widely used standardized questionnaire for the assessment of perceived usability (Klug, 2017; Lewis, 2018c). Sauro and Lewis (2009) reported that the SUS accounted for 43% of post-study questionnaire usage in a sample of industrial usability studies. Google Scholar citations (examined June 12, 2019) showed 8203 citations for the paper that introduced the SUS (Brooke, 1996). In its standard (most often used) form, the SUS has 10 five-point items with alternating positive and negative tone, and the following item content:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very awkward to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

Usability practitioners at Digital Equipment Corporation (DEC) developed the SUS in the mid-1980s (Dumas, 2003). The 10 five-point items of the SUS provide a unidimensional (no subscales) usability measurement that ranges from 0 to 100. In the first published account of the SUS, Brooke (1996) stated that the SUS was robust, reliable, and valid but did not publish any reliability or validity measurements. With regard to validity, "it correlates well with other subjective

measures of usability (e.g., the general usability subscale of the SUMI)" (Brooke, 1996, p. 194). According to Brooke (1996, p. 194), "the only prerequisite for its use is that any published report should acknowledge the source of the measure."

Since its initial publication, research on the SUS has led to some proposed changes in the original wording of the items. Finstad (2006) and Bangor et al. (2008) recommend the use of the word "awkward" rather than "cumbersome" in item 8. The original SUS items refer to "system. but substituting the word "product" or the use of the actual product name in place of "system" seems to have no effect on SUS scores (Lewis & Sauro, 2009), but, of course, substitutions should be consistent across the items.

To use the SUS, present the items to participants as five-point scales with response options numbered from 1 (anchored with "Strongly disagree") to 5 (anchored with "Strongly agree"). If a participant fails to respond to an item, assign it a 3 (the center of the rating scale). After completion, determine each item's score contribution, which will range from 0 to 4. For positively worded items (1, 3, 5, 7, and 9), the score contribution is the scale position minus 1. For negatively worded items (2, 4, 6, 8, and 10), it is 5 minus the scale position. To get the overall SUS score, multiply the sum of the item score contributions by 2.5. Thus, SUS scores range from 0 to 100 in 2.5-point increments.

Psychometric Evaluation

An early assessment of the SUS indicated reliability (assessed using coefficient alpha) of 0.85 (Lucey, 1991). More recent estimates indicate the reliability of the SUS is somewhat higher, typically around 0.90. For SUS translated into other languages, estimates of reliability tend to be a bit lower, though still acceptable, typically around 0.81 (Lewis, 2018c).

In addition to being highly reliable, recent studies have shown evidence of the validity of the SUS. Estimates of concurrent validity have ranged from correlations of 0.22 (with success rates, Peres, Pham, & Phillips, 2013) to 0.96 (with the UMUX, see Section 3.9.6, Finstad, 2010). Other findings include significant concurrent correlation with ratings of user friendliness (Bangor et al., 2008, r = 0.81), an adjective rating scale (Bangor, Kortum, & Miller, 2009, r = 0.50-0.79), likelihood-to-recommend (Lewis, Utesch, & Maher, 2015, r = 0.63), other standardized usability questionnaires (Berkman & Karahoca, 2016, r = 0.74; Lewis, 2018b, r = 0.74-0.79; Sauro, 2015b, r = 0.75), and success rates (Kortum & Peres, 2014, r = 0.73; Lah & Lewis, 2016, r = 0.50; Lewis, Utesch, & Maher, 2015, r = 0.50-0.63; Sauro, 2012, r = 0.90). Regarding its construct validity, the SUS, despite some evidence of bidimensionality published about ten years ago (Borsci, Federici, & Lauriola, 2009; Lewis & Sauro, 2009), appears to be unidimensional for all practical purposes, with the apparent bidimensionality likely due to its mixed tone (Lewis & Sauro, 2017b).

Tullis and Stetson (2004) provided early evidence of SUS sensitivity when they found that of five methods for assessing satisfaction with usability, the SUS was the quickest to converge on the "correct" conclusion regarding the usability of two websites as a function of sample size, where "correct" meant a significant *t*-test consistent with the decision reached using the total sample size. Later evidence of sensitivity includes sensitivity to product differences (Bangor et al., 2008; Finstad, 2010; Kortum & Bangor, 2013; Kortum & Sorber, 2015; Lewis, 2018b; Lewis & Sauro, 2009), personality types (Kortum & Oswald, 2017), prediction of business indicators (Bangor et al., 2013), and amount of experience (Kortum & Bangor, 2013; Lah & Lewis, 2016; McLellan et al., 2012).

Norms

One of the most useful aspects of the SUS is the availability of published norms for the interpretation of its scores. Starting in 2008, a number of UX researchers have collected large samples of SUS questionnaires and used them to develop interpretative norms. The first of these (Bangor et al., 2008) presented findings from almost 10 years of using the SUS in the evaluation of a large number of products in various stages of development (over 200 studies and more than 2300 completed SUS questionnaires). Some of their key findings were:

- The mean across all individual questionnaires was about 70, as was the mean computed across studies.
- Individual SUS scores ranged from 0 to 100, but across studies, the range of the means was more restricted, with 6% lower than a score of 50 and none lower than 30.
- Individual scores had a negative skew, but the distribution of study means was more normal.
- Inter-item correlations were consistently significant, ranging from 0.34 to 0.69.
- The SUS had an acceptable level of reliability (coefficient alpha of 0.91).
- The 10 items of the SUS all appeared to load on a single underlying factor.
- Comparison of six different classes of interface types (cell phones, customer equipment, graphical user interface, interactive voice response, Web, and Internet-based Web/IVR) found significant differences in SUS ratings as a function of interface type, which is evidence of scale sensitivity.
- There was evidence of a slight but significant negative relationship between score and age.
- There was no significant difference between male and female scores.

- Changes in SUS scores tracked logically with critical events in the product lifecycle process in a case study of iterative testing.

Given the large amount of SUS data collected over a decade, Bangor et al. (2008) made two attempts at developing norms with their data. About 10% (212) of the completed SUS questionnaires included an 11th item, an adjective rating scale with seven response options:

1: Worst imaginable (n = 1)
2: Awful (n = 0)
3: Poor (n = 15)
4: OK (n = 36)
5: Good (n = 90)
6: Excellent (n = 69)
7: Best imaginable (n = 1).

The SUS means for responses from 3–6 (for which $n \geq 15$) were, respectively after rounding to the nearest point, 39, 52, 73, and 86. The second approach was an absolute grading scale with A: 90–100, B: 80–89, C: 70–79, D: 60–69, and F: < 60.

Bangor et al. (2009) increased the sample size of concurrent collection of SUS with the adjective rating scale to almost 1000 cases. They reported a large and statistically significant correlation of 0.82 between the SUS and the adjective rating scale (evidence of concurrent validity). The means (and parenthetical sample sizes) for the seven response options were:

1: Worst imaginable = 12.5 (n = 4)
2: Awful = 20.3 (n = 22)
3: Poor = 35.7 (n = 72)
4: OK = 50.9 (n = 211)
5: Good = 71.4 (n = 345)
6: Excellent = 85.5 (n = 289)
7: Best imaginable = 90.9 (n = 16)

Note that Bangor et al. (2009) expressed some reservation over the interpretation of "OK" (with an associated mean SUS of 50.9) as suggesting an acceptable experience given an overall mean SUS closer to 70 in their large-sample data (Bangor et al., 2008). "In fact, some project team members have taken a score of OK to mean that the usability of the product is satisfactory and no improvements are needed, when scores within the OK range were clearly deficient in terms of perceived usability" (Bangor et al., 2009, p. 120). Their current practice is to

anchor this response option with “Fair” instead of “OK” (Phil Kortum, personal communication, February 22, 2018).

This line of research inspired the development of a curved rather than an absolute grading scale for the SUS (Sauro, 2011; Sauro & Lewis, 2012, 2016). Bangor et al. generously shared their SUS data with Jeff Sauro, as did Tullis and Albert (2013). With this combined data set from 446 studies and over 5000 individual SUS responses, Sauro (2011) used a logarithmic transformation on reflected scores to normalize the distribution, then computed percentile ranks for the entire range of SUS scores. Sauro and Lewis (2012, 2016) used those percentile ranks to create the curved grading scale (CGS) shown in Table 3.

To support quantitative analysis at the grade level, Table 3 includes a column of grade points based on the College Board method (College Board, 2019), suitable for computing a grade point average (GPA). When the focus is on detecting differences in SUS scores large enough have a perceptible effect on UX and sample sizes are large, statistically significant differences in GPA are more likely to represent practically significant effects than small but statistically significant differences in SUS scores.

Table 3 The Sauro–Lewis Curved Grading Scale for Interpreting the SUS

SUS score range	Grade	Grade point	Percentile range
84.1–100	A+	4.0	96–100
80.8–84.0	A	4.0	90–95
78.9–80.7	A-	3.7	85–89
77.2–78.8	B+	3.3	80–84
74.1–77.1	B	3.0	70–79
72.6–74.0	B-	2.7	65–69
71.1–72.5	C+	2.3	60–64
65.0–71.0	C	2.0	41–59
62.7–64.9	C-	1.7	35–40
51.7–62.6	D	1.0	15–34
0.0–51.6	F	0.0	0–14

Note that the average score in the data used to create the Sauro–Lewis CGS was 68, which was by design the exact center of the CGS (a grade of C), but would have been a D in the absolute grading scale. With its 11 grade categories, the CGS also provides a finer-grained scale than the adjective scale with its seven response options. It addresses the weakness of “OK” in the adjective scale because a 50 would receive an F (clearly deficient) while the lowest value in the range for C (an average experience) is 65. Finally, the CGS is consistent with an industrial practice that has become increasingly common of interpreting a mean SUS of at least 80 (A-) as indicative of an above average user experience. Throughout the rest of this chapter, letter grades are from the Sauro-Lewis CGS.

The Sauro–Lewis CGS provides good general guidance for the interpretation of SUS means. Several lines of research have shown, however, that different types of products and interfaces differ significantly in perceived usability. For example, Sauro (2011) partitioned his data from 446 studies into groups based on product type. The means (with associated CGS grades and number of studies) for some of the key categories were:

- Business-to-business software: 67.6 (C, n = 30)
- Mass market consumer software: 74.0 (B- n = 19)
- Public facing websites: 67.0 (C, n = 174)
- Internal productivity software: 76.7 (B, n = 21)

Kortum and Bangor (2013) published SUS ratings of overall experience for a set of 14 everyday products from a survey of more than 1000 users. Examples of the SUS means (with associated CGS grades and number of respondents) for products with low, medium, and high perceived usability were:

- Excel: 56.5 (D, n = 866)
- Word: 76.2 (B, n = 968)
- Amazon: 81.8 (A, n = 801)
- Google search: 92.7 (A+, n = 948)

There are different ways to interpret what these findings (Kortum & Bangor, 2013; Kortum & Sorber, 2015; Sauro, 2011) mean for industrial practice in user experience engineering. They could be interpreted as diminishing the value of the more general norms embodied in the CGS, but a more pragmatic interpretation is that they enhance the general norms. For example, consider the Kortum and Bangor ratings of everyday products. It should not be surprising that a complex spreadsheet program has lower perceived usability than a well-designed search box. For many projects, setting a SUS benchmark of 80 (A-) is reasonable and achievable. If, however, the project is to develop a competitive spreadsheet application, a SUS of 80 is probably unrealistically high (and is probably unrealistically low, if developing a new search interface). Where possible, practitioners should use a combination of comparison with norms and competitive evaluation when assessing the quality of their products. Practitioners should also exercise some caution when using data from within-subjects studies as benchmarks because respondents who are comparing products may, to a currently unknown extent, give slightly lower ratings to harder products and higher ratings to easier products than they otherwise might.

Item Benchmarks

Brooke (1996) cautioned against attempts to extract meaning from the items of the SUS, specifically, that the “SUS yields a single number representing a composite measure of the overall usability of the system being studied. Note that scores for individual items are not meaningful on their own” (p. 189). At the time, this

admonition was appropriate because his analyses were based on data from 20 people. With substantially more data in hand, Lewis and Sauro (2018) developed a series of regression equations for setting benchmarks for SUS items. This would not be useful for practitioners who are collecting data for attributes that are not one of the SUS items, such as findability. There are, however, some SUS items that might sometimes be useful apart from their contribution to the overall SUS, in particular, Item 2 (perceived complexity), Item 3 (perceived ease-of-use), Item 6 (perceived consistency), Items 7 or 10 (perceived learnability), and Item 9 (confidence in use). The data used to develop the regression equations came from 166 unpublished usability studies/surveys (a total of 11,855 individual SUS questionnaires). The 10 regression equations (with the text of the associated SUS item), computed using the means from the 166 individual studies, were:

- SUS01 = 1.073927 + 0.034024(SUS): "I think that I would like to use this system frequently."
- SUS02 = 5.834913—0.04980485(SUS): "I found the system unnecessarily complex."
- SUS03 = 0.4421485 + 0.04753406(SUS): "I thought the system was easy to use."
- SUS04 = 3.766087—0.02816776(SUS): "I think that I would need the support of a technical person to be able to use this system."
- SUS05 = 1.18663 + 0.03470129(SUS): "I found the various functions in this system were well integrated."
- SUS06 = 4.589912—0.03519522(SUS): "I thought there was too much inconsistency in this system."
- SUS07 = 0.9706981 + 0.04027653(SUS): "I would imagine that most people would learn to use this system very quickly."
- SUS08 = 5.575382—0.04896754(SUS): "I found the system very awkward to use."
- SUS09 = 0.6992487 + 0.04435754(SUS): "I felt very confident using the system."
- SUS10 = 4.603949—0.03692307(SUS): "I needed to learn a lot of things before I could get going with this system."

Note that due to the mixed tone of the SUS items the directionality of benchmarks would be different for odd and even-numbered items. For odd-numbered items, higher scores are better (using a basic five-point item scale); for even-numbered items lower scores indicate a better user experience. The first step in using the equations is to select a SUS value corresponding to a desired CGS grade level. For example, if a practitioner is interested in interpreting Item 3, "I thought the system was easy to use. then a mean score of 3.67 would correspond to a SUS mean of 68 (an average overall system score). For consistency with an above-average SUS mean of 80, the corresponding target for Item 3 would be an average score of at

least 4.24 (ideally statistically greater than the benchmark to control the risk of exceeding it by chance).

Flexibility of the SUS

Another of the strengths of the SUS in practical UX work is its flexibility, which extends beyond minor wording changes. In recent years, researchers have expanded its use beyond traditional usability testing to the retrospective measurement of perceived usability of products or classes of products (Grier et al., 2013; Kortum & Bangor, 2013). Grier et al. (2013) described a version of the SUS altered for the context of acquiring products for the U.S. military that are easy to troubleshoot and maintain, but did not provide any assessment of its psychometric properties in that context. Sauro and Lewis (2011) explored a more extreme manipulation of the SUS, specifically, changing the tone of the even-numbered items from negative to positive (2: "I found the system to be simple"; 4: I think that I could use the system without the support of a technical person"; 6: "I thought there was a lot of consistency in the system"; 8: "I found the system very intuitive"; 10: "I could use the system without having to learn anything new").

The original SUS was designed in accordance with a common strategy to control acquiescence bias, the hypothesized tendency of respondents to agree with statements, by having respondents rate statements with a mix of positive and negative tone. This practice also has potential benefits in helping researchers identify respondents who were not attentive to the statements they rated. There is, however, evidence that including a mix of positively and negatively worded items can create more problems than it solves (Barnette, 2000; Stewart & Frye, 2004), lowering internal reliability, distorting factor structure, and increasing interpretation problems in cross-cultural research. Furthermore, respondents may have difficulty switching response behaviors when completing questionnaires with mixed-tone items (mistakes), and researchers might forget the necessary step of reversing item scores for negative-tone items when computing overall scores (miscoding).

Sauro and Lewis (2011) administered a retrospective survey using the standard and positive versions of the SUS (n = 213) across seven websites. The reliability (coefficient alpha) of both questionnaires was high (Standard: 0.92; Positive: 0.96). The mean SUS scores for the two versions were not significantly different (Standard: 52.2; Positive: 49.3; $t(206) = 0.85$, $p > 0.39$). They found no evidence of acquiescence bias in either version, but estimated that about 17% of the completed standard questionnaires contained mistakes. They also reported that three of 27 SUS data sets (11%) contributed by anonymous donors to additional research efforts had miscoded SUS scores. "The data presented here suggest the problem of users making mistakes and researchers miscoding questionnaires is both real and much more detrimental than response biases" (Sauro & Lewis, 2011, p. 2221). Additional research using the positive version of the SUS has provided

evidence of its reliability, validity, and sensitivity (Lewis, Utesch, & Maher, 2013, 2015).

It is, however, possible to distort the SUS with items that have been rewritten to be unusually extreme. Sauro (2010b) described an experiment in which SUS items were manipulated to investigate two variables: item intensity and item tone. For example, the extreme negative version of the SUS Item 4 was "I think that I would need a permanent hot-line to the help desk to be able to use the website." The 62 participants were volunteers attending the 2008 conference of the Usability Professionals Association (UPA). They used one of five questionnaires to rate the UPA website: all positive extreme, all negative extreme, mixed extreme version 1, mixed extreme version 2, or the standard SUS. The scores from all positive extreme and all negative extreme were significantly different from the standard SUS.

Nine-Item Versions of the SUS

To compute the overall SUS score, respondents must provide a rating for each item. The instruction that Brooke (1996, p. 193) provided in the initial publication of the SUS was, "All items should be checked. If a respondent feels that they cannot respond to a particular item, they should mark the centre point of the scale." Thus, the typical practice when respondents do not provide a rating for an item is to replace the blank with the default rating of 3. But what if there is an item that would be confusing or distracting to respondents in a particular context of measurement? For example, the first SUS item is "I think I would like to use this system frequently." If the system under study is one that would only be used infrequently (e.g., a troubleshooting process or system for registering complaints), then there is a concern that including this item would distort the scores, or at best, distract the participant.

Lewis and Sauro (2017a) investigated the consequences of removing individual items from the standard SUS. Because previous research had indicated that small amounts of data missing from standardized usability questionnaires had little effect on the resulting scores (Lah & Lewis, 2016; Lewis, 2002) and the items of the SUS are significantly intercorrelated (Brooke, 1996), they hypothesized that the 10 possible nine-item versions of the SUS should not differ much from the score obtained with all 10 items given appropriate adjustment of the of the SUS multiplier.

To understand how to adjust the SUS multiplier, consider how the standard multiplier works. The process of determining score contributions described in the introduction results in a score that, without multiplication, would range from 0 to 40 (a maximum score contribution of 4 multiplied by 10 items). To stretch that out so it ranges from 0 to 100, it is necessary to multiply the sum of the score contributions by 100/40, which is the derivation of the "2.5" multiplier. After

dropping one item, the score contributions can range from 0 to 36 (9 × 4). To stretch this out to range from 0 to 100, the multiplier needs to be 100/36.

Lewis and Sauro (2017a) analyzed a data set of 9156 completed SUS questionnaires from 112 unpublished industrial usability studies and surveys. Note that with $n = 9156$, the study had the power to reliably detect very small differences and to precisely compute confidence intervals around estimated means, allowing a focus on differences that have practical rather than simply statistical significance (which only supports claims that differences are not plausibly 0). They computed the 10 possible nine-item scores that are possible when leaving one SUS item out, following the standard scheme for computing these SUS scores but multiplying the sum of the score contributions by 100/36 instead of 2.5 to compensate for the missing item. For each nine-item variant of the SUS, they assessed scale reliability using coefficient alpha, the correlation with the standard SUS, and the magnitude of the mean difference.

As expected, all nine-item variants of the SUS correlated significantly with the standard SUS (all $r > 0.99$). Dropping one item had no appreciable effect on scale reliability, with all values of coefficient alpha ranging from 0.90 to 0.91. The mean scores of all 10 possible nine-item variants of the SUS were within one point (out of 100) of the mean of the standard SUS. Thus, it appears that practitioners can leave out any one of the SUS items without having a practically significant effect on the resulting scores, as long as an appropriate adjustment is made to the multiplier (specifically, multiply the sum of the adjusted item scores by 100/36 instead of the standard 100/40, or 2.5, to compensate for the dropped item).

Translations

There have been a number of published translations of the SUS, including Arabic (AlGhannam, Albustan, Al-Hassan, & Albustan, 2017), Slovene (Blažica & Lewis, 2015), Polish (Borkowska & Jach, 2016), Italian (Borsci et al., 2009), Persian (Dianat, Ghanbari, & Asghari Jafarabadi, 2014), and Portuguese (Martinsa, Rosa, Queirós, & Silva, 2015).

The average estimate of reliability across these studies was about 0.81, lower than that typically found for the English version but well above the typical minimum criterion of 0.70. Estimates of concurrent validity with a variety of other metrics of perceived usability were significant correlations ranging from 0.45 to 0.95. Several studies found the SUS sensitive to the amount of experience with the product or system under investigation, consistent with sensitivity findings reported for the English version. Although there is no data currently available regarding its psychometric properties, a German version of the SUS is available (Rummel, 2015).

In Blažica and Lewis (2015), respondents rated the usability of the Slovene version of Gmail, providing an opportunity to compare those ratings with the

English assessment of Gmail reported in Kortum and Bangor (2013). The overall mean from the Slovene version was 81.7, close to the mean of 83.5 for Gmail reported by Kortum and Bangor (2013). Substantial overlap of the confidence intervals around these means indicated that the Gmail results for the Slovene version were reasonably close to the value published by Kortum and Bangor.

3.9.5 PSSUQ and CSUQ

The PSSUQ is a questionnaire designed for the purpose of assessing users' perceived satisfaction with their computer systems. It has its origin in an internal IBM project called SUMS (System Usability MetricS), headed by Suzanne Henry in the late 1980s (Lewis, 2019c). A team of human factors engineers and usability specialists working on SUMS created a pool of seven-point scale items based on the work of Whiteside et al. (1988) and from that pool selected 18 items to use in the first version of the PSSUQ (Lewis, 1992). Each item was worded positively, with the scale anchors "strongly agree" at the first scale position (1) and "strongly disagree" at the last scale position (7). A "not applicable" (NA) choice and a comment area were available for each item (see Lewis (1995) for examples of the appearance of the items). The questionnaire has been translated into Turkish (Erdinç & Lewis, 2013).

Psychometrics

Analyses of data from the SUMS project found the reliability of the overall summative scale (Overall) was 0.97, with acceptable subscale reliabilities (SysUse: 0.96, InfoQual: 0.91, IntQual: 0.91). The overall PSSUQ scores correlated highly with the sum of ASQ scores across the scenarios: $r(20) = 0.80$, $p < 0.0001$. The overall PSSUQ scores also correlated significantly with the percentage of successful scenario completions: $r(29) = -0.40$, $p = 0.026$. There was a highly significant correlation between SysUse and successful scenario completions: $r(36) = -0.40$, $p = 0.006$.

The development of the Computer System Usability Questionnaire (CSUQ) followed the development of the first version of the PSSUQ. Its items are identical to those of the PSSUQ except that their wording is appropriate for use in field settings or surveys rather than in a scenario-based usability test, making it, essentially, an alternative form of the PSSUQ. An unrelated series of IBM investigations into customer perception of usability revealed a common set of five usability characteristics associated with usability by several different user groups (Doug Antonelli, personal communication, January 5, 1991). The 18-item version of the PSSUQ addressed four of these five characteristics (quick completion of work, ease of learning, high-quality documentation and online information, and functional adequacy) but did not address the fifth (rapid acquisition of productivity).

The second version of the PSSUQ and CSUQ included an additional item to address this characteristic, bringing the total number of items up to 19. Analyses of data from a survey conducted with this version of the CSUQ (n = 377) confirmed the three-factor structure and replicated the reliability findings. The estimates of coefficient alpha for the CSUQ were 0.93 for SysUse, 0.91 for InfoQual, 0.89 for IntQual, and 0.95 for Overall (all within 0.03 of those from the initial PSSUQ study) (Lewis, 1995).

Lewis (2002) conducted a psychometric evaluation of the PSSUQ using data from several years of usability studies (primarily studies of speech dictation systems, but including studies of other types of applications). The results of a factor analysis on these data were consistent with earlier factor analyses (Lewis, 1992, 1995) used to define three PSSUQ subscales: system usefulness (SysUse), information quality (InfoQual), and interface quality (IntQual). Estimates of reliability were also consistent with those of earlier studies. Analyses of variance indicated that variables such as the specific study, developer, state of development, type of product, and type of evaluation significantly affected PSSUQ scores. Other variables, such as gender and completeness of responses to the questionnaire, did not. Norms derived from the new data correlated strongly with norms derived from earlier studies.

A potential criticism of the original PSSUQ has been that some items seemed redundant and that this redundancy might inflate estimates of reliability. Lewis (2002) investigated the effect of removing three items from the second version of the PSSUQ (items 3, 5, and 13). With these items removed, the reliability of the overall PSSUQ score (using coefficient α) was 0.94 (remaining very high), and the subscale reliabilities were:

- SysUse: 0.90
- InfoQual: 0.91
- IntQual: 0.83

All of the reliabilities exceeded 0.80, indicating sufficient reliability to be valuable as usability measurements (Anastasi, 1976; Landauer, 1997). Thus, the third (and current) version of the PSSUQ has 16 seven-point scale items (see Table 4 for the items and their normative scores).

Table 4 PSSUQ Version 3 Items, Scales, and Normative Scores

Item/scale	Item text/scale scoring rule[a]	Norm (99% CI)		
		Lower limit	Mean	Upper limit
Q1	Overall, I am satisfied with how easy it is to use this system.	2.60	2.85	3.09
Q2	It was simple to use this system.	2.45	2.69	2.93
Q3	I was able to complete the tasks and scenarios quickly using this system.	2.86	3.16	3.45
Q4	I felt comfortable using this system.	2.40	2.66	2.91
Q5	It was easy to learn to use this system.	2.07	2.27	2.48
Q6	I believe I could become productive quickly using this system.	2.54	2.86	3.17
Q7	The system gave error messages that clearly told me how to fix problems.	3.36	3.70	4.05
Q8	Whenever I made a mistake using the system, I could recover easily and quickly.	2.93	3.21	3.49
Q9	The information (such as on-line help, on-screen messages and other documentation) provided with this system was clear.	2.65	2.96	3.27
Q10	It was easy to find the information I needed.	2.79	3.09	3.38
Q11	The information was effective in helping me complete the tasks and scenarios.	2.46	2.74	3.01
Q12	The organization of information on the system screens was clear.	2.41	2.66	2.92
Q13	The interface[b] of this system was pleasant.	2.06	2.28	2.49
Q14	I liked using the interface of this system.	2.18	2.42	2.66
Q15	This system has all the functions and capabilities I expect it to have.	2.51	2.79	3.07
Q16	Overall, I am satisfied with this system.	2.55	2.82	3.09
SysUse	Average items 1–6.	2.57	2.80	3.02
InfoQual	Average items 7–12.	2.79	3.02	3.24
IntQual	Average items 13–15.	2.28	2.49	2.71
Overall	Average items 1–16.	2.62	2.82	3.02

a SysUse, system usefulness; InfoQual, information quality; IntQual, interface quality; CI, confidence interval. Scores can range from 1 (strongly agree) to 7 (strongly disagree), with lower scores better than higher scores.

b The "interface" includes those items that you use to interact with the system. For example, some components of the interface are the keyboard, the mouse, the microphone, and the screens (including their graphics and language).

Norms

Note that the scale construction is such that lower scores are better than higher scores and that the means of the items and scales all fall below the scale midpoint of 4. With the exception of item 7 ("The system gave error messages that clearly told me how to fix problems"), the upper limits of the confidence intervals are below 4. This shows that practitioners should not use the scale midpoint exclusively as a reference from which they would judge participants' perceptions of usability. Rather, they should also use the norms shown in Table 4 (and comparison with these norms is probably more meaningful than comparison with the scale midpoint).

The way that item 7 stands out from the others indicates:

- This pattern should not surprise practitioners if it is in their data.
- Providing usable error messages throughout a product is difficult.
- It may well be worth the effort to focus on providing usable error messages.
- Finding the mean for this item to be equal to or less than the mean of the other items in InfoQual (assuming they are in line with the norms), is an indication of success in creating better-than-average error messages.

The consistent pattern of relatively poor ratings for InfoQual versus IntQual [seen across all the studies; for details and complete normative data, see Lewis (2002, 2019)] suggests that practitioners who find this pattern in their data should not conclude that they have poor documentation or a great interface.

Another potential criticism of the PSSUQ is that the items do not follow the typical convention of varying the tone of the items so that half of the items elicit agreement and the other half elicit disagreement (Swamy, 2007). The rationale for the decision to align the items consistently was to make it as easy as possible for participants to complete the questionnaire. With consistent item alignment, the proper way to mark responses on the items is clearer, potentially reducing response errors due to participant confusion. Also, the use of negatively worded items can produce a number of undesirable effects (Barnette, 2000; Ibrahim, 2001; Sauro & Lewis, 2011), including problems with internal consistency and factor structure. Additional key findings and conclusions from Lewis (2002) were:

- There was no evidence of response styles (especially, no evidence of extreme response style) in the PSSUQ data.
- Because there is a possibility of extreme response and acquiescence response styles in cross-cultural research (Baumgartner & Steenkamp, 2001; Clarke, 2001; Grimm & Church, 1999; van de Vijver & Leung, 2001), practitioners should avoid using questionnaires for cross-cultural comparison unless that use has been validated. Other types of group comparisons with the PSSUQ are valid because any

effect of response style should cancel out across experimental conditions.

- Scale scores from incomplete PSSUQs were indistinguishable from those computed from complete PSSUQs. This data does not provide information concerning how many items a participant might ignore and still produce reliable scale scores. It does suggest that, in practice, participants typically complete enough items to produce reliable scale scores. The similarity of psychometric properties across the various versions of the PSSUQ, despite the passage of time and differences in the types of systems studied, provides evidence of significant generalizability for the questionnaire, supporting its use by practitioners for measuring participant satisfaction with the usability of tested systems. Due to its generalizability, practitioners can confidently use the PSSUQ when evaluating different types of products and at different times during the development process. The PSSUQ can be especially useful in competitive evaluations (for an example, see Lewis, 1996) or when tracking changes in usability as a function of design changes made during development. Practitioners and researchers are free to use the PSSUQ and CSUQ (no license fees), but anyone using them should cite the source.

Correspondence with SUS

There have been three studies published with concurrent collection of CSUQ and SUS data, enabling investigation of the extent to which CSUQ and SUS scores correspond in magnitude. Should they have substantial correspondence, then it would be possible for the PSSUQ and CSUQ to "piggy-back" on the norms developed for the SUS (Table 3).

To assess the correspondence between the means, it's helpful to convert the CSUQ to a metric that, like the SUS, can range from 0 to 100 where higher scores indicate a better user experience. The process of getting from a traditional CSUQ score to one that matches the SUS involves subtracting 1 from the mean of the 16 individual CSUQ items and multiplying that by 100/6 to stretch it out to a 0-100-point scale, then subtracting from 100 to reverse the scale. For example, if the mean CSUQ was 1 (the best possible standard CSUQ mean), the transformed score would be 100 (100 - (1 - 1)(100/6) = 100 - 0 = 100). If the mean CSUQ was 7 (the worst possible standard CSUQ mean), the transformed score would be 0 (100 - (7 - 1)(100/6) = 100 - 100 = 0). For a mean CSUQ of 4 (the center of the standard CSUQ 7-point scale), the transformed score would be 50 (100 - (4 - 1)(100/6) = 100 - 50 = 50).

Table 5 shows the SUS and CSUQ results from the three studies, with seven independent estimates of CSUQ/SUS correspondence, both for scores and grade point averages. Across the estimates, the mean of the difference scores was 1.6, with a 95% confidence interval ranging from 0.3 to 2.9. After translation to a grade point value using Table 3, the mean difference in GPA was 0.1, with a 95% confidence interval ranging from 0.0 to 0.3. These findings support CSUQ interpretation with SUS norms.

Table 5 CSUQ Correspondence with the SUS

Product (study)	SUS mean	CSUQ mean	Mean Diff	SUS CGS	CSUQ CGS	SUS GPA	CSUQ GPA	GPA Diff
Mind Maps (Berkman & Karahoca, 2016)	79.5	80.0	-0.5	A-	A-	3.7	3.7	0.0
Windows OS (Lewis, 2018b)	66.9	64.1	2.8	C	C-	2.0	1.7	0.3
Apple OS (Lewis, 2018b)	76.8	76.6	0.2	B	B	3.0	3.0	0.0
Excel (Lewis, 2019a)	69.6	68.7	0.9	C	C	2.0	2.0	0.0
Word (Lewis, 2019a)	75.5	72.8	2.7	B	B-	3.0	2.7	0.3
Amazon (Lewis, 2019a)	84.8	82.3	2.5	A+	A	4.0	4.0	0.0
Gmail (Lewis, 2019a)	78.0	75.3	2.7	B+	B	3.3	3.0	0.3

3.9.6 UMUX and UMUX-LITE

UMUX

There are some situations in which a shorter instrument is preferable to a longer one (e.g., when there is a need to measure more attributes than just perceived usability leading to limited "real estate" for any given attribute). The UMUX (Finstad, 2010, 2013) was designed at Intel to get a measurement of perceived usability consistent with the SUS, but using only the following four items (presented as 7-point agreement scales anchored on the left with "Strongly disagree" and on the right with "Strongly agree"):

- This system's capabilities meet my requirements.
- Using this system is a frustrating experience.
- This system is easy to use.
- I have to spend too much time correcting things with this system.

Like the standard SUS, UMUX item scores are manipulated to obtain an overall score that ranges from 0 to 100. In addition to the initial research by Finstad (2010), other researchers (Berkman & Karahoca, 2016; Borsci et al., 2015; Lewis, Utesch, & Maher, 2013, 2015) have also reported desirable psychometric properties for the UMUX, including acceptable levels of:

- reliability (coefficient alpha greater than 0.80);
- concurrent validity (correlation with SUS greater than 0.55; correlation with CSUQ equal to −0.65);
- sensitivity to different levels of a variety of independent variables (e.g., discriminating between systems of independently assessed levels of relatively good and poor usability, detecting differences in perceived usability as a function of experience).

Most research in this area has found substantial correspondence between the magnitudes of mean SUS and UMUX. An exception is Borsci et al. (2015), who reported UMUX means that were significantly and markedly higher than concurrently collected SUS means. Despite this, averaging across 13 estimates, the mean of the difference scores was -1.9, with a 95% confidence interval ranging from -4.9 to 1.1. After translation to a grade point value using Table 3, the mean difference in GPA was -0.3, with a 95% confidence interval ranging from -0.6 to 0.1. These findings support interpreting UMUX scores with SUS norms.

Analyses of the factor structure of the UMUX have been inconsistent. With only four items, the most likely structures are one or two factors. Finstad (2010, 2013) reported a one-factor structure. Lewis, Utesch, and Maher (2013, 2015) found a two-factor structure reflecting the positive/negative item tone. Berkman and Karahoca (2016) replicated the two-factor positive/negative tone structure when forcing a two-factor solution, but also reported evidence from confirmatory factor analysis suggesting a one-factor structure.

Following the same practical reasoning as that for the SUS, it doesn't matter whether the UMUX has a unidimensional or tone-based bidimensional structure—in either case, practitioners should treat the UMUX as a unidimensional measurement of perceived usability. Like the CSUQ and the SUS, the UMUX and measures derived from it are available for use by researchers or practitioners without a license fee.

UMUX-LITE

The UMUX-LITE (Lewis, Utesch, & Maher, 2013, 2015) is a short version of the UMUX consisting of its positive-tone items (selected based on factor and item analysis), which are:

1 This system's capabilities meet my requirements.
2 This system is easy to use.

There are two versions of the UMUX-LITE that usability practitioners should be aware of, and they should also be aware that the UMUX-LITE literature has been inconsistent in its terminology. The formula for computing the standard UMUXLITE, where x_1 and x_2 are the ratings for Items 1 and 2 using a standard 7-point scale (1–7), is: UMUXLITE = $(x_1 + x_2 - 2)(100/12)$. Due to a small but statistically significant difference between the SUS and UMUX-LITE means, Lewis et al. (2013) computed a regression equation to bring the SUS and UMUX-LITE scores into closer correspondence, naming that adjustment the UMUX-LITEr.

Because the UMUX-LITEr is a linear adjustment of the UMUX-LITE, it has many of the same statistical properties, such as the magnitude of correlation with other metrics, but can only take values between 22.9 (when UMUX-LITE = 0) and 87.9 (when UMUX-LITE = 100). This range restriction has the effect of diminishing UMUX-LITEr estimates when corresponding SUS means are above average (B+ or higher on the Sauro-Lewis curved grading scale), so Lewis (2019a, 2019c) has recommended using the standard UMUX-LITE in research and practice rather than the regression-adjusted UMUX-LITEr.

Practitioners should also be aware of some practitioner variation in the number of response options used in the UMUX-LITE (Sauro, 2017b). Lewis (2019b) reported that the number of UMUX-LITE response options did not matter much for 5-, 7-, or 11-response options, especially in practice, but recommended against using 3-response options due to some weakness with regard to reliability and correlation with likelihood-to-recommend.

In addition to the statistical analyses supporting their selection (Lewis et al., 2013), it is interesting that the content of the two items of the UMUX-LITE matches the constructs of the Technology Acceptance Model (TAM) (Davis, 1989), a questionnaire from the information systems literature that assesses the perceived usefulness (e.g., capabilities meeting requirements) and perceived ease-of-use of systems, and has an established relationship to likelihood of future use. According to the TAM, good ratings of perceived usefulness and ease of use (perceived usability) influence the intention to use, which in turn influence the actual likelihood of use.

Research on the UMUX-LITE (Berkman & Karahoca, 2016; Lah et al., 2020; Lewis, 2018b, 2019a; Lewis et al., 2013, 2015) has demonstrated acceptable psychometric properties, including:

- acceptable reliability (estimates of coefficient alpha ranging from 0.76 to 0.86)
- concurrent validity (correlations with SUS ranging from 0.74 to 0.86; correlation with ratings of likelihood-to-recommend ranging from 0.72 to 0.74)

- sensitivity (significant differences as a function of respondents' ratings of frequency-of-use)
- on average, close correspondence with concurrently collected SUS data.

Table 6 shows the correspondence between concurrently collected SUS and UMUX-LITE means. Across the 13 estimates, the average difference between the SUS and UMUX-LITE means was -0.6—less than one point for metrics that can range from 0 to 100. The 95% confidence interval around the estimate ranged from -2.4 to 1.3. For the grade point averages, the mean difference was -0.1, with a 95% confidence interval ranging from -0.4 to 0.2.

Table 6 UMUX-LITE Correspondence with the SUS

Product (study)	SUS mean	UMUX -LITE mean	Mean Diff	SUS CGS	UMUX -LITE CGS	SUS GPA	UMUX -LITE GPA	GPA Diff
Mind Maps (Berkman & Karahoca, 2016)	79.5	78.5	1.0	A-	B+	3.7	3.3	0.4
PowerPoint (Lah et al., 2020)	70.8	74.3	-3.5	C	B	2.0	3.0	-1.0
Gmail (Lah et al., 2020)	79.3	81.2	-1.9	B+	A	3.7	4.0	-0.3
Notes (Lah et al., 2020)	56.8	59.3	-2.5	D	D	1.0	1.0	0.0
Apple OS (Lewis, 2018b)	76.8	79.9	-3.1	B	A-	3.0	3.7	-0.7
Windows OS (Lewis, 2018b)	66.9	68.5	-1.6	C	C	2.0	2.0	0.0
Excel (Lewis, 2019a)	69.6	74.0	-4.4	C	B-	2.0	2.7	-0.7
Word (Lewis, 2019a)	75.5	78.0	-2.5	B	B+	3.0	3.3	-0.3
Amazon (Lewis, 2019a)	84.8	86.6	-1.8	A+	A+	4.0	4.0	0.0
Gmail (Lewis, 2019a)	78.0	77.7	0.3	B+	B+	3.3	3.3	0.0
Various (Lewis et al., 2013)	53.5	50.3	3.2	D	F	1.0	0.0	1.0
Various (Lewis et al., 2013)	58.8	55.1	3.7	D	D	1.0	1.0	0.0
Various (Lewis, Utesch, & Maher, 2015)	58.1	52.4	5.7	D	D	1.0	1.0	0.0

SUS and UMUX-LITE measures appear to be reasonably consistent when considering mean raw differences and are very consistent for mean grade point differences. These findings support the use of the UMUX-LITE as a concise UX metric that can be interpreted using the Sauro–Lewis curved grading scale (Table 3).

There also appears to be a connection between UMUX-LITE and a commonly used business measure of customer loyalty, the Net Promoter Score (NPS, Reichheld, 2003, 2006). Friedman and Flaounas (2018) reported, for their specific context of measurement, a substantial correlation between UMUX-LITE and NPS (r = 0.62), and an associated regression formula for the relationship, $NPS = 3.18(UMUXLITE) - 200.6$, which led them to state (p. 603):

> For example, a business goal of increasing the NPS of the product by 20 points from 20 to 40, translates to a goal of increasing the UMUX-LITE score from 69.37 [CGS Grade of C] to 75.66 [CGS Grade of B]. The particular coefficients may change across different products, companies or periods – NPS can be affected by factors other than usability, and different products and companies will have different UMUX-LITE and NPS baselines. However, a similar approach can be applied to assess the relation between the metrics in a particular context, and inform the product teams when they derive their goals from higher-level business goals.

3.9.7 Other User Experience Questionnaires

Sauro and Lewis (2016) devoted an entire chapter to standardized usability questionnaires, including five instruments for quick post-task assessment and 22 longer questionnaires designed to capture a variety of usability/UX measures (such as the questionnaires discussed above: QUIS, SUMI, SUPR-Q, PSSUQ, CSUQ, UMUX and UMUX-LITE). This section provides an introduction to post-task questionnaires and a few relatively new UX questionnaires.

Post-task Questionnaires

Questionnaires designed for use after a usability study or even longer retrospective times in surveys are important tools for UX researchers and practitioners, but their measurements are at a relatively high level. For this reason, UX researchers and practitioners often perform a quick assessment of perceived usability immediately after participants complete each task in a usability study. Research indicates a correlation of 0.64 between post-study and post-task assessments of perceived usability (Sauro & Lewis, 2009), supporting the practice of taking both types of

measurements when conducting studies. Some of the approaches to post-task measurement are:

- *After-Scenario Questionnaire (ASQ):* This is a three-item questionnaire developed at the same time and using the same format as the PSSUQ (Lewis, 1991b, 1995). The items address ratings of ease of task completion, satisfaction with task completion time, and satisfaction with supporting information. Its reported reliability ranges from 0.90 to 0.96, and it significantly correlates with successful task completion.
- *Single Ease Question (SEQ):* The SEQ is one of the most widely used post-task questionnaires. It has one item, usually seven response options with endpoints of "Very difficult" on the left and "Very easy" on the right. Despite differences in format, its content is very similar to the first item of the ASQ and the second item of the UMUX-LITE. Researchers have reported significant correlations between the SEQ and objective usability (Tedesco & Tullis, 2006, efficiency; Sauro & Dumas, 2009, completion times, number of errors), as well as with other measures of perceived usability (Sauro & Dumas, 2009, SUS).
- *Subjective Mental Effort Question (SMEQ):* Zijlstra and van Doorn (1985) published the SMEQ, a single vertical item that takes ratings from 0 to 150, anchored at the bottom with "Not at all hard to do" just above 0, at just above 110 with "Tremendously hard to do. and with seven other labels in between. SMEQ correlates significantly with concurrently collected SEQ and SUS measurements, as well as with completion times and number of errors (Sauro & Dumas, 2009).
- *Expectation Ratings (ER):* The ER method requires pre-task assessment of how easy participants think a task will be, followed by post-task assessment of their experience with the task (Albert & Dixon, 2003), using items similar to the SEQ. This requires additional effort, but allows for the graphing of results into four quadrants: Promote It (anticipated to be difficult, but was easy), Big Opportunity (anticipated to be difficult, and it was, so room for improvement), Don't Touch It (perceived as easy before and after task completion), and Fix It Fast (anticipated to be easy but turned out to be difficult).
- *Usability Magnitude Estimation (UME):* Magnitude estimation has its roots in psychophysics, the branch of

> psychology that explores mathematical relationships between the physical properties of a stimulus and its perception. There have been a few attempts to apply the technique of magnitude estimation to perceived usability (Cordes, 1984a, 1984b; McGee, 2003, 2004), with mixed success (Sauro & Dumas, 2009; Tedesco & Tullis, 2006). Even after training, participants seem to have trouble with concepts like "twice as difficult."

Tedesco and Tullis (2006) compared three 5-point variants of the SEQ, the ASQ (first two items only), and ER. Using a subsampling procedure and varying sample sizes from 3 to 29, they found the standard version of the SEQ to be most sensitive. Sauro and Dumas (2009) conducted a similar resampling experiment with the recommended (Sauro & Lewis, 2016) 7-point version of the SEQ, the SMEQ, and UME, concluding that the SEQ and SMEQ were more sensitive than UME.

PUTQ

The Purdue Usability Testing Questionnaire (Lin, Choong, & Salvendy, 1997) was developed from consideration of eight human factors relevant to software usability—compatibility, consistency, flexibility, learnability, minimal action, minimal memory load, perceptual limitation, and user guidance. In a small-sample experiment, it correlated highly with concurrently collected QUIS scores (concurrent validity) and was more sensitive than the QUIS in discriminating the usability of two interface designs.

EMO

The Emotional Metric Outcomes (EMO) questionnaire (Lewis, Brown, & Mayes, 2015; Lewis & Mayes, 2014) was designed to assess the emotional outcomes of interaction, especially the interaction of customers with service-provider personnel or software. It is available in 16- and 8-item versions, with subscales for Positive Relationship Affect, Negative Relationship Affect, Positive Personal Affect, and Negative Personal Affect. Across three surveys, its overall reliability was high (around 0.94), and subscale reliabilities ranged from 0.76 to 0.94. Factor and regression analysis generally confirmed the expected four-factor structure, and it has been shown to be sensitive to industry differences and successful task completion.

AttrakDiff2

The AttrakDiff2 questionnaire (Diefenbach et al., 2014; Hassenzahl, 2018) is based on research conducted over almost 20 years, starting with Hassenzahl's psychometric distinction between pragmatic quality (traditional perceived usability) and hedonic quality (emotional aspects of use) and demonstration how

both factors contribute to the appeal of a product (Hassenzahl et al., 2000; Hassenzahl, 2001). The current version of AttrakDiff consists of 28 7-point semantic differential items (e.g., “confusing-clear” for pragmatic quality; “unusual-ordinary” for hedonic), providing measures of pragmatic quality and two aspects of hedonic quality, Stimulation (novelty, challenge) and Identification (self-expression) (Hassenzahl, 2004). An 8-item short version is also available (Hassenzahl & Monk, 2010). As an example of its use in UX research, Hassenzahl et al. (2015) used AttrakDiff2 to explore aspects of experience-oriented and product-oriented evaluation.

UEQ

Another questionnaire influenced by the work of Hassenzahl is the User Experience Questionnaire (UEQ; Laugwitz, Held, & Schrepp, 2008; Rauschenberger et al., 2013). Like AttrakDiff2, the UEQ assesses pragmatic and hedonic quality with 7-point semantic differential items, but across 26 items has subdivisions of pragmatic quality into Perspicuity, Efficiency, and Dependability and subdivisions of hedonic quality into Novelty and Stimulation. Concurrently collected AttrakDiff2 and UEQ data have shown the expected convergent and divergent correlations, and all UEQ scale reliabilities were acceptable (Laugwitz, Schrepp, & Held, 2008). The UEQ is available in 17 languages (Rauschenberger et al., 2012; Schrepp, Hinderks, & Thomaschewski, 2017a), and in an 8-item short version (Schrepp, Hinderks, & Thomaschewski, 2017b).

Schrepp et al. (2017a) published benchmarks for the UEQ based on evaluations of 246 products (100 complex business applications, 4 development tools, 64 web shops or services, 3 social networks, 16 mobile applications, 20 household appliances, and 39 other products). Across these evaluations, there were 9905 responses, with responses per evaluation ranging from 3 to 1390. Most evaluations were of mature products, commercially developed and designed. The UEQ convention for scoring is to set 0 at the middle of the 7-point scale, making the minimum possible score -3 and the maximum possible score +3. For this set of data, the means (and standard deviations) for each of the benchmarks were:

- Attractiveness: 1.04 (0.64)
- Efficiency: 0.97 (0.62)
- Perspicuity: 1.06 (0.67)
- Dependability: 1.07 (0.52)
- Stimulation: 0.87 (0.63)
- Originality: 0.61 (0.72)

In addition to these means, Schrepp et al. (2017a) published finer-grained benchmarks for each scale based on product percentiles in their normative database (best 10% = Excellent, 10% better and 75% worse = Good, 25% better

and 50% worse: Above Average, 50% better and 25% worse = Below Average, worst 25% = Bad). For the full set of benchmarks, see their Table 1. For example:

- *Excellent:* Att. ≥ 1.75; Eff. ≥ 1.78; Per. ≥ 1.9; Dep. ≥ 1.65; Sti. ≥ 1.55; Nov. ≥ 1.4
- *Above Average:* 1.17 ≥ Att. < 1.52; 0.98 ≥ Eff. < 1.47; 1.08 ≥ Per. < 1.56; 1.14 ≥ Dep. < 1.48; 0.99 ≥ Sti. < 1.31; 0.71 ≥ Nov. < 1.05
- *Bad:* Att. < 0.7; Eff. < 0.54; Per. < 0.64; Dep. < 0.78; Sti. < 0.5; Nov. < 0.3

Schrepp et al. (2017a) recommended collecting data from 20–30 users to get a stable measurement. They did not discuss a method for computing an overall score from the UEQ scales, instead recommending that new products achieve at least Good ratings on all scales. They also suggested that practitioners give some thought to which scales might be the most important for their product's context of use, and strive to achieve Excellent for those scales. Finally, they expressed a desire to improve their benchmarks by distinguishing among different types of products in future research, once sufficient data are available.

meCUE

The meCUE is a standardized questionnaire based on the Components of User Experience (CUE) model published by Thüring and Mahlke (2007). In that model, the components are the perception of non-instrumental product qualities such as aesthetics, status, and commitment; emotions, and perception of instrumental qualities such as perceived usefulness and perceived usability. The meCUE (Minge, Thüring, Wagner, & Kuhr, 2016) was designed to assess each of these components in a modular fashion ("me" is short for "modular evaluation"), using 7-point Likert-style agreement scales (all points labeled).

The initial pool of 67 items came from brainstorming sessions and inspection of existing questionnaires. Data collected in two surveys, each with n = 238, were used to guide item selection. Principal components analysis indicated retention of five components for instrumental and non-instrumental qualities, named Usefulness, Usability, Visual Aesthetics, Status, and Commitment, all combined in the first module. For each component in this module, the three items with the highest component loadings were included. For the second module (Emotions) the process of item selection was more complex, resulting in the inclusion of items with a mix of positive/negative valence and high/low arousal. Items for product loyalty and intention to use made up the third module. The resulting questionnaire had 33 items measuring nine dimensions clustered into three modules.

To test its internal consistency and validity, 67 participants completed typical tasks with three different interactive products and completed a set of

questionnaires including AttrakDiff, UEQ, PANAS (Positive and Negative Affect Scales, Watson et al., 1988), Self-Assessment Manikin (Bradley & Lang, 1994), a visual aesthetics questionnaire (Lavie & Tractinsky, 2004), and the first version of meCUE. The results indicated reliable scales that were consistent with the expected factor structures and correlated as expected with the other questionnaires. There were also significant correlations between the number of completed tasks and the meCUE Usefulness, Usability, and Product Loyalty scales. Following this analysis, an additional item was added to assess the overall experience (as in the AttrakDiff and UEQ's Attractiveness metric). A second experiment indicated appropriate discriminative and convergent validities for ratings of experiences with applications differing in usability and aesthetics.

After using the meCUE for a few years and after reanalysis of the data from Minge et al. (2016), Minge and Thüring (2018) separated the first module into two separate modules, one for instrumental and the other for non-instrumental qualities, naming this revised version the meCUE 2.0. The questionnaire is available in its original German and an English translation.

3.9.8 Guidance on Which Usability/UX Questionnaire(s) to Use

For after-task questionnaires, the most common is SEQ because it is one simple item that directly assesses perceived ease-of-use. The SMEQ also has good measurement properties, but is a bit more complex and can be more difficult for certain types of respondents to use for whom dragging a slider is more difficult than clicking a radio button. If additional measurement of satisfaction with completion time and supporting materials is important, use the ASQ. If additional measurement of perceived ease-of-use before and after task completion is important, use ER. Due to its difficulty of use in practical situations, we do not recommend using UME.

There are a number of factors to consider when selecting which longer usability/UX questionnaire to use, summarized in Table 7 (norms, length, scales, and fees).

Table 7 Characteristics of Key Standardized Usability/UX Questionnaires

Questionnaire	Norms	# Items	# Scales	Scale labels	Fees
QUIS	No	26/71	12	Overall Reaction, Screen Factors, Terminology and System Feedback, Learning Factors, System Capabilities, Technical Manuals, Multimedia, Voice Recognition, Virtual Environments, Internet Access, Software Installation	Yes
SUMI	Curated	50	6	Global, Efficiency, Affect, Helpfulness, Control, Learnability	Yes
SUPR-Q	Curated	8	5	Global, Usability, Trust, Loyalty, Appearance	Yes
SUPR-Qm	Curated	16	5	Global, Usability, Trust, Loyalty, Appearance	No
SUS	Public	10	1	Perceived Usability	No
PSSUQ/CSUQ	Public	16	4	Overall, System Quality, Information Quality, Interface Quality	No
UMUX	Public	4	1	Perceived Usability	No
UMUX-LITE	Public	2	1	Perceived Usability	No
EMO	No	16/8	5	Overall, Positive Relationship Affect, Negative Relationship Affect, Positive Personal Affect, Negative Personal Affect	No
AttrakDiff2	No	28/8	3	Pragmatic Quality, Stimulation, Identification	Yes
UEQ	Public	26/8	5	Perspicuity, Efficiency, Dependability, Novelty, Stimulation	No
meCUE	No	33	5	Overall, Instrumental Qualities, Non-Instrumental Qualities, Emotions, Loyalty	No

Norms

Will it be necessary to benchmark questionnaire scores against a set of norms? If so, and if using public open-source norms is acceptable, the SUS and UMUX-LITE are probably the best unidimensional choices. The PSSUQ/CSUQ and UEQ are very different multidimensional questionnaires, both in the content of their items and in whether they provide a way to compute an overall score from their subscales (yes for PSSUQ/CSUQ; no for UEQ). If higher-quality curated norms are important, consider using the SUMI or SUPR-Q. This will be the case when it is necessary to use norms that are kept up-to-date and are specific to different product types or industries.

Length

The length of the various questionnaires ranges from 2 (UMUX-LITE) to 71 (long version of QUIS) items. In general, choose the shortest questionnaire that otherwise meets the measurement goals of the study. Note that several questionnaires are available in long and short forms (QUIS, EMO, AttrakDiff2, UEQ)—usually the longer forms will provide more reliable measurement, but there are many questionnaires that have acceptable (often more than acceptable) reliability with just two items per scale in their short forms.

Scales

Choose questionnaires that have measurement scales appropriate for the study. Of the questionnaires reviewed in this section, three (SUS, UMUX, UMUX-LITE) produce overall scores for just one scale—perceived usability (although the first UMUX-LITE item provides an assessment of perceived usefulness). The number of scales measured in the other questionnaires ranges from 3 to 12, and vary substantially in the aspects of UX they intend to measure.

Fees

Most of these questionnaires are available for use without a fee, but there are some exceptions, especially for questionnaires that have been developed and are curated by companies or academic institutions. If planning to use the QUIS, SUMI, SUPR-Q, SUPR-Qm, or AttrakDiff2, it is important to review their conditions of use, which are available on their websites or by contacting their owners.

4 WRAPPING UP

4.1 Getting More Information about Usability and UX Design and Evaluation

This chapter has provided information about usability and UX design and evaluation, but there is only so much that you can cover in a single chapter. For additional chapter-length treatments of the basics of usability testing, see Nielsen (1997), Dumas (2003), and Dumas and Salzman (2006). The classic books devoted to the topic of usability testing are Dumas and Redish (1999), Rubin (1994); Rubin and Chisnell, (2008), and Barnum (2002). Although their references and examples may be a bit dated, they all deserve a place on the usability/UX practitioner's bookshelf because in many ways the practice of usability testing has not changed, and the assessment of usability is central to the assessment of UX.

Krug's (2009) *Rocket Surgery Made Easy* is a popular introduction to low-investment usability testing. For a more comprehensive but still practical treatment, see Barnum's (2011) *Usability Testing Essentials: Ready, Set…Test*!

Tullis and Albert's (2013) *Measuring the User Experience: Collecting, Analyzing and Presenting Usability Metrics* is a book-length treatment of user experience measurement, with a companion website at `www.measuringux.com`. Sauro and Lewis' *Quantifying the User Experience* (2016) is a book-length treatment of statistical methods for usability testing and other user research applications.

Relatively recent book-length treatments of user experience design and evaluation include Allanwood and Beare (2019) *User Experience Design: A Practical Introduction*, Still and Crane (2017) *Fundamentals of User-Centered Design*, and Nunnally and Farkas (2017) *UX Research*.

You can also get information on usability/UX design and research from the magazines and journals produced by professional organizations such as ACM, HFES, and UXPA, for example:

- *Ergonomics in Design*
- *Interactions*
- *Journal of Usability Studies*
- *Human Factors*
- *International Journal of Human-Computer Interaction*

- *IEEE Transactions on Professional Communication*
- *IEEE Software*
- *Behaviour & Information Technology*
- *Behavior Research Methods*
- *Communications of the ACM*
- *Applied Ergonomics*
- *Computers in Human Behavior*
- *Interacting with Computers*
- *International Journal of Human-Computer Studies*

For late-breaking developments in usability and UX research and practice, there are a number of annual conferences that have these topics as significant portions of their content. Companies making a sincere effort in the professional development of their usability practitioners should ensure that their personnel have access to the proceedings of these conferences and should support attendance at one or more of these conferences at least every few years. These major conferences are:

- User Experience Professionals Association (www.uxpa.org)
- Human–Computer Interaction International (www.hci-international.org)
- ACM Special Interest Group in Computer–Human Interaction (www.acm.org/sigchi)
- Human Factors and Ergonomics Society (hfes.org)
- INTERACT (held every two years; see, e.g., www.interact2019.org)

4.2 Usability/UX Design and Evaluation: Yesterday, Today, and Tomorrow

It seems clear that iterative design and usability testing (both summative and formative) are here to stay and that their general form will remain similar to the forms that emerged in the late 1970s and early 1980s. The last 40 years have seen the introduction of a variety of usability/UX evaluation techniques and some consensus (and some continuing debate) on the conditions under which to use the various techniques, either alone or in combination (Al-Wabil & Al-Khalifa, 2009; Hornbæk, 2010; Jarrett et al., 2009). In the last 30 years, usability researchers have made significant progress in the areas of standardized usability questionnaires and sample size estimation for formative usability tests.

In the past 20 years there have been significant advances in large-sample remote usability testing (Albert et al., 2010; Sauro, 2018b). Given its emerging focus on commercial self-service, there has been additional research in and development of standardized usability questionnaires for the Internet (Bargas-Avila, Lötscher, Orsini, & Opwis, 2009; Joyce & Kirakowski, 2015; Lascu & Clow, 2008, 2013), with extensions to address Internet-specific factors such as

trust and other elements of customer experience from the marketing research literature (Lewis & Mayes, 2014; Safar & Turner, 2005).

In the last 10 years there has been substantial research in the think-aloud method (e.g., Alhadreti & Mayhew, 2017, 2018; Elabour, Alhadreti, & Mayhew, 2017; Hertzum et al., 2015; Hertzum & Holmegaard, 2013, 2015; Karahasanovic et al., 2009; McDonald, Edwards, & Zhao, 2012; McDonald, McGarry, & Willis, 2013; McDonald, Zhao, & Edwards, 2013) and research in the relationship between UX and business goals (e.g., Bangor et al., 2013; Friedman & Flaounas, 2018; Oliveira et al., 2017), connections among UX metrics (e.g., Berkman & Karahoca, 2016; Borsci et al., 2015; Lewis, 2018b, 2019a; Lewis & Sauro, 2017a, 2017b, 2018; Lewis et al., 2013, 2015), and expansion of the scope of UX in standardized questionnaires (e.g., Diefenbach et al., 2014; Hassenzahl, 2018; Kortum & Bangor, 2013; Lewis & Mayes, 2014; Minge & Thüring, 2018; Sauro, 2015; Sauro & Zarolia, 2017; Schrepp et al., 2017a, 2017b).

As we look to the future, we can anticipate additional research on topics such as improved modeling of problem discovery and appropriate sample size for formative research (e.g., Hertzum et al., 2014; Hwang & Salvendy, 2010; Lewis, 2014; Schmettow, 2012), more studies to untangle the outcome differences due to variations of TA testing; improved modeling of the relationship among components of UX and potential antecedents and consequences in research and business (e.g., Alonso-Rios et al., 2010; Borsci et al., 2018; Grishin & Gillan, 2019; Lewis, 2018b; Sonderegger & Sauer, 2010; Tractinsky, 2017; Tuch et al., 2012), development of additional UX scales and continuing work on comparative scaling of UX questionnaires with the SUS grading scale (e.g., Lewis, 2019c), and more translations of UX questionnaires to support UX research, design, and evaluation across the world (e.g., Blažica & Lewis, 2015; Erdinç & Lewis, 2013; Schrepp et al., 2017a). Finally, there would be significant value in replicating the research of Bailey (1993) which has provided compelling evidence supporting the effectiveness of iterative design and usability testing, but which is only one study, now over 25 years old.

In the meantime, practitioners will continue to perform iterative usability/UX design and evaluation, exercising professional judgment as required. For example, usability testing is not a perfect usability evaluation method in the sense that it does not guarantee the discovery of all possible usability problems, but it does not have to be perfect to be useful and effective. It is, however, important to understand the strengths, limitations, and current leading practices to ensure proper (most effective) use of usability/UX design and evaluation methods.

REFERENCES

Abelson, R. P. (1995). *Statistics as principled argument*. Mahwah, NJ. Lawrence Erlbaum Associates.

Al-Awar, J., Chapanis, A., & Ford, R. (1981). Tutorials for the first-time computer user. *IEEE Transactions on Professional Communication*, *24*, 30–37.

Albert, W., & Dixon, E. (2003). Is this what you expected?: The use of expectation measures in usability testing. Paper presented at the Usability Professionals Association Annual Conference, UPA, Scottsdale, AZ.

Albert, W., Tullis, T., & Tedesco, D. (2010). *Beyond the usability lab: Conducting large-scale user experience studies*. Burlington, MA: Morgan-Kaufmann.

AlGhannam, B. A., Albustan, S. A., Al-Hassan, A. A., & Albustan, L. A. (2017). Towards a standard Arabic System Usability Scale (A-SUS): Psychometric evaluation using communication disorder app. *International Journal of Human-Computer Interaction*, *34*(9), 799–804.

Alhadreti, O. & Mayhew, P. (2017). To intervene or not to intervene: An investigation of three think-aloud protocols in usability testing. *Journal of Usability Studies, 12*(3), 111–132.

Alhadreti, O. & Mayhew, P. (2018). Are two pairs of eyes better than one? A comparison of concurrent think-aloud and co-participation methods in usability testing. *Journal of Usability Studies, 13*(4), 177–195.

Allanwood, G. & Beare, P. (2019). *User experience design: A practical introduction* (2nd ed.). London: Bloomsbury Visual Arts.

Alonso-Ríos, D., Vázquez-Garcia, A., Mosqueira-Rey, E., & Moret-Bonillo, P. (2010). usability: A critical analysis and a taxonomy. *International Journal of Human-Computer Interaction*, *26*(1), 53–74.

Al-Wabil, A. & Al-Khalifa, H. (2009). A framework for integrating usability methods: The Mawhiba web portal case study. In *Proceedings of Current Trends in Information Technology 2009* (pp. 1–6). IEEE, Dubai, UAE.

Anastasi, A. (1976). *Psychological testing*. New York: Macmillan.

Andre, T. S., Belz, S. M, McCreary, F. A., & Hartson, H. R. (2000). Testing a framework for reliable classification of usability problems. In *Proceedings of the IEA 2000/HFES 2000 Congress*, Human Factors and Ergonomics Society, Santa Monica, CA, pp. 573–576.

Andreasen, M. S., Nielsen, H. V., Schrøder, S. O., & Stage, J. (2007). What happened to remote usability testing? An empirical study of three methods. In *Proceedings of CHI 2007* (pp. 1405–1414). Association for Computing Machinery, San Jose, CA.

ANSI (American National Standards Institute) (2001). *Common industry format for usability test reports*, ANSI-NCITS 354-2001. Washington, DC: ANSI.

Aranyi, G., & van Schaik, P. (2015). Modeling user experience with news websites. *Journal of the Association for Information Science and Technology, 66*(12), 2471–2493.

Arning, K., & Ziefle, M. (2008). Development and validation of a computer expertise questionnaire for older adults. *Behaviour & Information Technology, 27*(1), 89–93.

Aykin, N. M. & Aykin, T. (1991). Individual differences in human–computer interaction. *Computers and Industrial Engineering, 20*, 373–379.

Baecker, R. M. (2008). Themes in the early history of HCI: Some unanswered questions. *Interactions, 15*(2), 22–27.

Bailey, G. (1993). Iterative methodology and designer training in human–computer interface design. In *INTERCHI '93 Conference Proceedings* (pp. 198–205). Association for Computing Machinery, New York.

Bailey, R. W., Allan, R. W., & Raiello, P. (1992). Usability testing vs. heuristic evaluation: A head to head comparison. In *Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting* (pp. 409–413). Human Factors and Ergonomics Society, Santa Monica, CA.

Bailey, R. W., Wolfson, C. A., Nall, J., & Koyani, S. (2009). Performance-based usability testing: Metrics that have the greatest impact for improving a system's usability. In M. Kurosu (Ed.), *Human centered design, HCII 2009* (pp. 3–12). Springer-Verlag, Heidelberg, Germany.

Balentine, B. & Morgan, D. P. (2001). *How to build a speech recognition application: a style guide for telephony dialogues* (2nd ed.). San Ramon, CA: EIG Press.

Ballard, B. (2007). *Designing the mobile experience*. Chichester: Wiley.

Bangor, A., Joseph, K., Sweeney-Dillon, M., Stettler, G., & Pratt, J. (2013). Using the SUS to help demonstrate usability's value to business goals. In *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting* (pp. 202–205). HFES, Santa Monica, CA.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human-Computer Interaction, 24*, 574–594.

Bangor, A., Kortum, P. T., & Miller, J. T. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies, 4*, 114–123.

Barbaroux, M. (2016). *Untangling UX, Part 1: Design thinking vs. UCD*. Downloaded 8/23/19 from `https://www.cambridgeconsultants.com/insights/untangling-ux-part-1-design-thinking-vs-ucd`.

Bargas-Avila, J. A. & Hornbæk, K. (2011). Old wine in new bottles or novel challenges? A critical analysis of empirical studies of user Experience. In *Proceedings of CHI*

2011 (pp. 2689–2698). Association for Computing Machinery, Vancouver, Canada.

Bargas-Avila, J. A., Lötscher, J., Orsini, S., & Opwis, K. (2009). Intranet satisfaction questionnaire: Development and validation of a questionnaire to measure user satisfaction with the intranet. *Computers in Human Behavior, 25,* 1241–1250.

Barnette, J. J. (2000). Effects of stem and likert response option reversals on survey internal consistency: if you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement, 60,* 361–370.

Barnum, C. (2002). *Usability testing and research*. New York: Longman.

Barnum, C. (2011). *Usability testing essentials: ready, set…test*! Burlington, MA: Morgan Kaufmann.

Baumgartner, H. & Steenkamp, J. B. E. M. (2001). Response styles in marketing research: a cross-national investigation. *Journal of Marketing Research, 38,* 143–156.

Bennett, J. L. (1979). The commercial impact of usability in interactive systems. *Infotech State of the Art Report: Man/Computer Communication, 2,* 289–297.

Berkman, M. I. & Karahoca, D. (2016). Re-Assessing the Usability Metric for User Experience (UMUX) scale. *Journal of Usability Studies, 11*(3), 89–109.

Berry, D. C. & Broadbent, D. E. (1990). The role of instruction and verbalization in improving performance on complex search tasks. *Behaviour & Information Technology, 9,* 175–190.

Bevan, N. (2009). Extending quality in use to provide a framework for usability measurement. In M. Kurosu (Ed.), *Human Centered Design, HCII 2009* (pp. 13–22). Springer-Verlag, Heidelberg, Germany.

Bevan, N., Kirakowski, J., & Maissel, J. (1991). What is usability? In H. J. Bullinger (Ed.), *Human Aspects in Computing, Design and Use of Interactive Systems and Work with Terminals, Proceedings of the 4th International Conference on Human–Computer Interaction* (pp. 651–655). Elsevier Science, Stuttgart, Germany.

Bias, R. G. & Mayhew, D. J. (1994). *Cost-justifying usability*. Boston: Academic.

Billestrup, J., Bruun, A., & Stage, J. (2016). Usability problems experienced by different groups of skilled internet users: Gender, age, and background. In *Proceedings of IFIP 2016* (pp. 45–55). Springer International Publishing, Stockholm, Sweden.

Bitner, M. J., Ostrom, A. L., & Meuter, M. L. (2002). Implementing successful self-service technologies. *Academy of Management Executive, 16*(4), 96–108.

Blalock, H. M. (1972). *Social statistics*. New York: McGraw-Hill.

Blažica, B. & Lewis, J. R. (2015). Slovene translation of the System Usability Scale: The SUS-SI. *International Journal of Human-Computer Interaction, 31*(2), 112–117.

Boren, T. & Ramey, J. (2000). Thinking aloud: reconciling theory and practice. *IEEE Transactions on Professional Communications, 43*, 261–278.

Borkowska, A, & Jach, K. (2016). "Pre-testing of Polish translation of System Usability Scale (SUS). In J. Świątek, Z. Wilimowska, L., Borzemski, & A. Grzech (Eds.) *Proceedings of 37th International Conference on Information Systems Architecture and Technology—ISAT 2016—Part I* (pp. 143–153). Springer, New York.

Borsci, S., Federici, S., Bacci, S., Gnaldi, M., & Bartolucci, F. (2015). Assessing user satisfaction in the era of user experience: comparison of the SUS, UMUX, and UMUX-LITE as a function of product experience. *International Journal of Human-Computer Interaction, 31*, 484–495.

Borsci, S., Federici, S., & Lauriola, M. (2009). On the dimensionality of the system usability scale: A test of alternative measurement models. *Cognitive Processes, 10*, 193–197.

Borsci, S., Federici, S., Malizia, A., & de Filippis, M. L. (2018). Shaking the usability tree: Why usability is not a dead end, and a constructive way forward. *Behaviour & Information Technology*, DOI: 10.1080/0144929X.2018.1541255.

Borsci, S., MacRedie, R. D., Barnett, J., Martin, J., Kuljis, J., & Young, T. (2013). Reviewing and extending the five-user assumption: A grounded procedure for interaction evaluation. *ACM Transactions on Computer-Human Interaction, 20*(5), Article 29.

Bosenick, T., Kehr, S., Kühn, M., & Nufer, S. (2007). Remote usability tests: An extension of the usability toolbox for online-shops. In C. Stephanidis (Ed.), *Universal access in HCI, Part I, HCII 2007* (pp. 392–398). Springer-Verlag, Heidelberg, Germany.

Bowers, V. & Snyder, H. (1990). Concurrent versus retrospective verbal protocols for comparing window usability. In *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 1270–1274). Human Factors Society, Santa Monica, CA.

Bradley, M. M. & Lang, P. J. (1994). Measuring emotions: the self-assessment manikin and the semantic differential. *Journal of Behavioral Therapy and Experimental Psychiatry*, *25*(1), 49–59.

Brooke, J. (1996). SUS: A 'quick and dirty' usability scale. In P. Jordan, B. Thomas, & B. Weerdmeester (Eds.), *Usability evaluation in industry* (pp. 189–194). London: Taylor & Francis.

Brooke, J. (2013). SUS: A retrospective. *Journal of Usability Studies, 8*(2), 29–40.

Capra, M. G. (2007). Comparing usability problem identification and description by practitioners and students. In *Proceedings of the Human Factors and Ergonomics Society 51st Annual Meeting* (pp. 474–478). Human Factors and Ergonomics Society, Santa Monica, CA.

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction*, Lawrence Erlbaum, Hillsdale, NJ.

Caulton, D. A. (2001). Relaxing the homogeneity assumption in usability testing. *Behaviour & Information Technology, 20*, 1–7.

Cavallin, H., Martin, W. M., & Heylighen, A. (2007). How relative absolute can be: SUMI and the impact of the nature of the task in measuring perceived software usability. *Artificial Intelligence and Society, 22*, 227–235.

Chapanis, A. (1981). Evaluating ease of use. unpublished manuscript prepared for IBM, Boca Raton, FL, available from J. R. Lewis.

Chapanis, A. (1988). Some generalizations about generalization. *Human Factors, 30*, 253–267.

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988. Development of an instrument measuring user satisfaction of the human–computer Interface. In E. Soloway, D. Frye, & S. B. Sheppard (Eds.), *CHI '88 Conference Proceedings: Human Factors in Computing Systems* (pp. 213–218). Association for Computing Machinery, Washington, DC.

Clarke, I. (2001). Extreme response style in cross-cultural research. *International Marketing Review, 18*, 301–324.

Clemmensen, T., Hertzum, M., Hornbæk, K., Shi, Q., & Yammiyavar, P. (2009). Cultural cognition in usability evaluation. *Interacting with Computers, 21*, 212–220.

Cliff, N. (1987) *Analyzing multivariate data*. San Diego, CA: Harcourt Brace Jovanovich.

Cockton, G. & Lavery, D. (1999). A framework for usability problem extraction. In M. A. Sasse & C. Johnson (Eds.), *Human Computer Interaction, INTERACT '99* (pp. 344–352). Amsterdam: IOS Press.

Cockton, G., Lavery, D., & Woolrych, A. (2003. Inspection-based evaluations. In J. A. Jacko & A. Sears (Eds.), *The human-computer interaction handbook* (pp. 1118–1138). Mahwah, NJ: Lawrence Erlbaum.

College Board. (2019). How to convert your GPA to a 4.0 scale. Downloaded on 10/12/2019.from `https://pages.collegeboard.org/how-to-convert-gpa-4.0-scale`

Cordes, R. E. (1984a). Software ease of use evaluation using magnitude estimation. In *Proceedings of the Human Factors Society* (pp. 157–160). HFS, Santa Monica, CA.

Cordes, R. E. (1984b). Use of magnitude estimation for evaluating product ease of use (Tech. Report 82.0135) .Tucson, AZ: IBM.

Cordes, R. E. (1993). The effects of running fewer subjects on time-on-task measures. *International Journal of Human–Computer Interaction, 5*, 393–403.

Cordes, R. E. (2001). Task-selection bias: A case for user-defined tasks. *International Journal of Human–Computer Interaction, 13*, 411–419.

Curedale, R. (2019). *Design thinking: Process and methods* (5th ed.). Topanga, CA: Design Community College.

Davis, D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–339.

Dianat, I., Ghanbari, Z., & AsghariJafarabadi, M. (2014). Psychometric properties of the persian language version of the system usability scale. *Health Promotion Perspectives, 4*(1), 82–89.

Dickens, J. (1987). The fresh cream cakes market: The use of qualitative research as part of a consumer research programme. In U. Bradley (Ed.), *Applied marketing and social research* (pp. 23–68). New York: Wiley.

Diefenbach, S., Kolb, N., & Hassenzahl, M. (2014). The 'hedonic' in human-computer interaction—history, contributions, and future research directions. In *Proceedings of DIS 2014* (pp. 305–314). Association for Computing Machinery, Vancouver, BC.

Dow, S., MacIntyre, B., Lee, J., Oczbek, C., Bolter, J. D., & Gandy, M. (2005). Wizard of Oz support throughout an iterative design process. *Pervasive Computing, 4*(4), 18–26.

Downey, L. L. (2007). Group usability testing: Evolution in usability techniques. *Journal of Usability Studies, 2*(3), 133–144.

Dumas, J. S. (2003). User-based evaluations. In J. A. Jacko & A. Sears (Eds.), *The human–computer interaction handbook* (pp. 1093–1117). Mahwah, NJ: Lawrence Erlbaum Associates.

Dumas, J. (2007). The great leap forward: The birth of the usability profession (1988–1993). *Journal of Usability Studies, 2*(2), 54–60.

Dumas, J. & Redish, J. C. (1999). *A practical guide to usability testing*. Portland, OR: Intellect.

Dumas, J. & Salzman, M. C. (2006). Usability assessment methods. In R. C. Williges (Ed.), *Reviews of human factors and ergonomics* (vol. 2, pp. 109–140). Santa Monica, CA: Human Factors and Ergonomics Society.

Dumas, J., Sorce, J., & Virzi, R. (1995). Expert reviews: how many experts is enough? In *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting* (pp. 228–232). Human Factors and Ergonomics Society, Santa Monica, CA.

Elbabour, F., Alhadreti, O., & Mayhew, P. (2017). Eye tracking in retrospective think-aloud usability testing: is there added value? *Journal of Usability Studies, 12*(3), 95–110.

Erdinç, O., & Lewis, J. R. (2013). Psychometric evaluation of the T-CSUQ: The Turkish version of the computer system usability questionnaire. *International Journal of Human-Computer Interaction, 29*, 319–323.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*, 215–251.

Finstad, K. (2006). The system usability scale and non-native English speakers. *Journal of Usability Studies, 1*(4), 185–188.

Finstad, K. (2010). The usability metric for user experience. *Interacting with Computers, 22*, 323–327.

Finstad, K. (2013). Response to commentaries on 'The usability metric for user experience'. *Interacting with Computers, 25*, 327–330.

Fisher, J. (1991). Defining the novice user. *Behaviour & Information Technology, 10*, 437–441.

Følstad, A., Law, E., & Hornbæk, K. (2012). Analysis in practical usability evaluation: A survey study. In *Proceedings of CHI 2012* (pp. 2127–2136). ACM, Austin, TX.

Fowler, C. J. H., Macaulay, L. A., & Fowler, J. F. (1985). The relationship between cognitive style and dialogue style: An exploratory study. In P. Johnson & S. Cook (Eds.), *People and computers: Designing the interface* (pp. 186–198). Cambridge: Cambridge University Press.

Friedman, A., & Flaounas, I. (2018). The right metric for the right stakeholder: a case study of improving product usability. In *Proceedings of OzCHI 2018* (pp. 602–606). Association for Computing Machinery, Melbourne, Australia.

Gawron, V. J., Drury, C. G., Czaja, S. J., & Wilkins, D. M. (1989). A taxonomy of independent variables affecting human performance. *International Journal of Man–Machine Studies, 31*, 643–672.

Genov, A., Keavney, M., & Zazelenchuk, T. (2009). Usability testing with real data. *Journal of Usability Studies, 4*(2), 85–92.

Gillan, D. & Schvaneveldt, R. W. (1999). Applying cognitive psychology: Bridging the gulf between basic research and cognitive artifacts. In *Handbook of applied cognition* (pp. 3–31). New York: Wiley.

Gordon, W. & Langmaid, R. (1988). *Qualitative market research: A practitioner's and buyer's guide*, Aldershot: Gower.

Gould, J. D. (1988). How to design usable systems. In M. Helander (Ed.), *Handbook of human–computer interaction* (pp. 757–789). North.-Holland: Amsterdam.

Gould, J. D. & Boies, S. J. (1983). Human factors challenges in creating a principal support office system: the speech filing system approach. *ACM Transactions on Information Systems, 1*, 273–298.

Gould, J. D., Boies, S. J., Levy, S., Richards, J. T., & Schoonard, J. (1987). The 1984 olympic message system: A test of behavioral principles of system design. *Communications of the ACM, 30*, 758–769.

Gould, J. D. & Lewis, C. (1984). *Designing for usability: Key principles and what designers think*, Technical Report RC-10317, Yorktown Heights, NY: International Business Machines Corporation.

Gray, W. D. & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human–Computer Interaction, 13*, 203–261.

Greene, S. L., Gomez, L. M., & Devlin, S. J. (1986). A cognitive analysis of database query production. In *Proceedings of the 30th Annual Meeting of the Human Factors Society* (pp. 9–13). Human Factors Society, Santa Monica, CA.

Grier, R. A., Bangor, A., Kortum, P., & Peres, S. C. (2013). The system usability scale: beyond standard usability testing. In *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting* (pp. 187–191). HFES, Santa Monica, CA.

Grimm, S. D. & Church, A. T. (1999). A cross-cultural study of response biases in personality measures. *Journal of Research in Personality, 33*, 415–441.

Grishin, J.,& Gillan, D. J. (2019). Exploring the boundary conditions of the effect of aesthetics on perceived usability. *Journal of Usability Studies, 14*(2), 76–104.

Guan, Z., Lee, S., Cuddihy, E., & Ramey, J. (2006). The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In *Proceedings of CHI 2006* (pp. 1253–1262). Association for Computing Machinery, Montreal, Quebec.

Gulliksen, J., Göransson, B., Boivie, I., Blomkvist, S., Persson, J., & Cajander, Å. (2003). Key principles for user-centered system design. *Behaviour & Information Technology, 22*(6), 397–409.

Hackman, G. S. & Biers, D. W. (1992). Team usability testing: Are two heads better than one? In *Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting* (pp. 1205–1209). Human Factors and Ergonomics Society, Santa Monica, CA.

Hassenzahl, M. (2000). Prioritizing usability problems: data driven and judgement driven severity estimates. *Behaviour & Information Technology, 19*, 29–42.

Hassenzahl, M. (2001). The effect of perceived hedonic quality on product appealingness. *International Journal of Human-Computer Interaction, 13*(4), 481–499.

Hassenzahl, M. (2004). The interplay of beauty, goodness, and usability in interactive products. *Human-Computer Interaction, 19*, 319–349.

Hassenzahl, M. (2018). A personal journey through user experience. *Journal of Usability Studies, 13*(4), 168–176.

Hassenzahl, M. & Monk, A. (2010). The inference of perceived usability from beauty. *Human-Computer Interaction, 25*, 235–260.

Hassenzahl, M., Platz, A., Burmester, M., & Lehner, K. (2000). Hedonic and ergonomic quality aspects determine a software's appeal. In *Proceedings of CHI 2000* (pp. 201–208). Association for Computing Machinery, The Hague, The Netherlands,

Hassenzahl, M., & Tractinsky, N. (2006). User experience—a research agenda. *Behaviour & Information Technology, 25*(2), 91–97.

Hassenzahl, M., Wiklund-Engblom, A., Bengs, A., Hägglund, S., & Diefenbach, S. (2015). Experience-oriented and product-oriented evaluation: psychological need

fulfillment, positive affect, and product perception. *International Journal of Human-Computer Interaction, 31*, 530–544.

Hertzum, M. (2006). Problem prioritization in usability evaluation: from severity assessments to impact on design. *International Journal of Human-Computer Interaction, 21*(2), 125–146.

Hertzum, M. (2010). Images of usability. *International Journal of Human-Computer Interaction, 26*(6), 567–600.

Hertzum, M., Borlund, P., & Kristoffersen, K. B. (2015). What do thinking-aloud participants say? A comparison of moderated and unmoderated usability sessions. *International Journal of Human-Computer Interaction, 31,* 557–570.

Hertzum, M. & Clemmensen, T. (2012). How do usability professionals construe usability? *International Journal of Human-Computer Studies, 70,* 26–42.

Hertzum, M., Hansen, K. D., & Andersen, H. H. K. (2009). scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology, 28*(2), 165–181.

Hertzum, M. & Holmegaard, K. D. (2013). Thinking aloud in the presence of interruptions and time constraints. *International Journal of Human-Computer Interaction, 29,* 351–364.

Hertzum, M.,& Holmegaard, K. D. (2015). Thinking aloud influences perceived time. *Human Factors, 57*(1), 101–109.

Hertzum, M. & Jacobsen, N. J. (2003). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human–Computer Interaction, 15,* 183–204.

Hertzum, M., Molich, R., & Jacobsen, N. E. (2014). What you get is what you see: Revisiting the evaluator effect in usability tests. *Behaviour & Information Technology, 33*(2), 143–161.

Høegh, R. T. & Jensen, J. J. (2008). A case study of three software projects: Can software developers anticipate the usability problems in their software? *Behaviour & Information Technology, 27*(4), 307–312.

Høegh, R. T., Nielsen, C. M., Overgaard, M., Pedersen, M. B., & Stage, J. (2006). The Impact of usability reports and user test observations on developers' understanding of usability data: An exploratory study. *International Journal of Human-Computer Interaction, 21*(2), 173–196.

Holleran, P. A. (1991). A methodological note on pitfalls in usability testing. *Behaviour & Information Technology, 10,* 345–357.

Hornbæk, K. (2006). Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies, 64*(2), 79–102.

Hornbæk, K. (2010). Dogmas in the assessment of usability evaluation methods. *Behaviour & Information Technology, 29*(1), 97–111.

Hornbæk, K. & Frøkjær, E. (2008a). A study of the evaluator effect in usability testing. *Human-Computer Interaction, 23*, 251–277.

Hornbæk, K. & Frøkjær, E. (2008b). Comparison of techniques for matching of usability problem descriptions. *Interacting with Computers, 20*, 505–514.

Hornbæk, K. & Law, E. L. (2007). Meta-analysis of correlations among usability measures. In *Proceedings of CHI 2007* (pp. 617–626). Association for Computing Machinery, San Jose, CA.

Howard, T. (2008). Unexpected complexity in a traditional usability study. *Journal of Usability Studies, 3*(4), 189–205.

Howard, T. & Howard, W. (2009). Unexpected complexity in user testing of information products. In *Proceedings of the Professional Communication Conference* (pp. 1–5). Institute of Electrical and Electronics Engineers, Waikiki, HI.

Hwang, W. & Salvendy, G. (2007). What makes evaluators to find more usability problems?: A meta-analysis for individual detection rates. In J. Jacko (Ed.), *Human-Computer Interaction, Part I, HCII 2007* (pp. 499–507). Springer-Verlag, Heidelberg, Germany.

Hwang, W. & Salvendy, G. (2009). Integration of usability evaluation studies via a novel meta-analytic approach: What are significant attributes for effective evaluation. *International Journal of Human-Computer Interaction, 25*(4), 282–306.

Hwang, W. & Salvendy, G. (2010). Number of people required for usability evaluation: The 10+2 Rule. *Communications of the ACM, 53*(5), 130–133.

Ibrahim, A. M. (2001). Differential responding to positive and negative items: the case of a negative item in a questionnaire for course and faculty evaluation. *Psychological Reports, 88*, 497–500.

Illmensee, T. & Muff, A. (2009). 5 users every Friday: A case study in applied research. In *Proceedings of the 2009 Agile Conference* (pp. 404–409). Institute of Electrical and Electronics Engineers, Chicago, IL.

International Organization for Standardization (ISO) (1998). *Ergonomic requirements for office work with visual display terminals (VDTs)* Part 11, *Guidance on usability*, ISO 9241-11:1998(E), Geneva, Switzerland: ISO.

Jarrett, C. & Gaffney, G. (2009). *Forms that work: Designing web forms for usability*, Burlington, MA: Morgan Kaufmann.

Jarrett, C., Quesenbery, W., Roddis, I., Allen, S., & Stirling, V. (2009). Using measurements from usability testing, search log analysis and web traffic analysis to inform development of a complex web site used for complex tasks. In M. Kurosu (Ed.), *Human Centered Design, HCII 2009* (pp. 729–738). Springer-Verlag, Heidelberg, Germany.

Jokela, T., Koivumaa, J., Pirkola, J., Salminen, P., & Kantola, N. (2006). Methods for quantitative usability requirements: a case study on the development of the user interface of a mobile phone. *Personal and Ubiquitous Computing, 10*, 345–355.

Jordan, P. (2002.) *Designing pleasurable products: An introduction to the new human factors*. London: CRC Press.

Joyce, M. & Kirakowski, J. (2015). Measuring attitudes towards the internet: the general attitude scale. *International Journal of Human-Computer Interaction, 31*, 506–517.

Karahasanovic, A., Hinkel, U. N., Sjøberg, D. I. K., & Thomas, R. (2009). Comparing of feedback-collection and think-aloud methods in program comprehension studies. *Behaviour & Information Technology, 28*(2), 139–164.

Karat, C. (1997). Cost-justifying usability engineering in the software life cycle. In M. Helander, T. K. Landauer, & P. Prabhu (Eds.), *Handbook of human–computer interaction* (2nd ed., pp. 767–778). Amsterdam: Elsevier.

Karat, J. (1997). User-centered software evaluation methodologies. In M. Helander, T. K. Landauer, & P. Prabhu (Eds.), *Handbook of human–computer interaction* (2nd ed., pp. 689–704). Amsterdam: Elsevier.

Karat, J. & Karat, C. (2003). The evolution of user-centered focus in the human-computer interaction field. *IBM Systems Journal, 42*(4), 532–541.

Keenan, S. L., Hartson, H. R., Kafura, D. G., & Schulman, R. S. (1999). The usability problem taxonomy: A framework for classification and analysis. *Empirical Software Engineering, 1,* 71–104.

Kelley, J. F. (1984). An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems, 2*, 26–41.

Kelley, J. F. (1985). CAL: A natural language program developed with the OZ paradigm: implications for supercomputing systems. In *Proceedings of the First International Conference on Supercomputing Systems* (pp. 238–248). Association for Computing Machinery, New York.

Kelley, J. F. (2018). Wizard of Oz (WoZ)—A yellow brick journey. *Journal of Usability Studies, 13*(3), 119–124.

Kennedy, P. J. (1982). Development and testing of the operator training package for a small computer system. In *Proceedings of the Human Factors Society 26th Annual Meeting* (pp. 715–717). Human Factors Society, Santa Monica, CA.

Kessner, M., Wood, J., Dillon, R. F., & West, R. L. (2001). On the reliability of usability testing. In J. Jacko & A. Sears (Eds.), *Conference on Human Factors in Computing Systems: CHI 2001 extended abstracts* (pp. 97–98). Association for Computing Machinery, Seattle, WA.

Kirakowski, J. (1996). The Software Usability Measurement Inventory: Background and usage. In P. Jordan, B. Thomas, & B. Weerdmeester (Eds.), *Usability evaluation in industry* (pp. 169–178). London: Taylor & Francis.

Kirakowski, J. & Corbett, M. (1993). SUMI: The Software Usability Measurement Inventory. *British Journal of Educational Technology, 24*, 210–212.

Kirakowski, J. & Dillon, A. (1988). The Computer User Satisfaction Inventory (CUSI): manual and scoring key. Human Factors Research Group, University College of Cork, Cork, Ireland.

Klug, B. (2017). An overview of the System Usability Scale in library website and system usability testing. *Weave: Journal of Library User Experience, 1*(6), 1–20.

Kortum, P. & Bangor, A. (2013). Usability ratings for everyday products measured with the System Usability Scale. *International Journal of Human-Computer Interaction, 29*, 67–76.

Kortum, P. & Johnson, M. (2013). The relationship between levels of user experience with a product and perceived system usability. In *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting* (pp. 197–201). HFES, Santa Monica, CA.

Kortum, P. & Oswald, F. L. (2017). The impact of personality on the subjective assessment of usability. *International Journal of Human-Computer Interaction, 34*(2), 177–186.

Kortum, P. & Peres, S. C. (2014). The relationship between system effectiveness and subjective usability scores using the System Usability Scale. *International Journal of Human-Computer Interaction, 30*, 575–584.

Kortum, P. & Sorber, M. (2015). Measuring the usability of mobile applications for phones and tablets. *International Journal of Human-Computer Interaction, 31*(8), 518–529.

Krahmer, E. & Ummelen, N. (2004). Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Transactions on Professional Communication, 47*(2), 105–117.

Krug, S. (2009). *Rocket surgery made easy: The do-it-yourself guide to finding and fixing usability problems*. Berkeley, CA: New Riders.

Krug, S. (2014). *Don't make me think, revisited: A common sense approach to web and mobile usability* (3rd ed.). Berkeley, CA: New Riders.

Lah, U. & Lewis, J. R. (2016). How expertise affects a digital-rights-management-sharing application's usability. *IEEE Software, 33*(3), 76–82.

Lah, U., Lewis, J. R., & Šumak, B. (2020). Perceived usability and the modified technology acceptance model. *International Journal of Human-Computer Interaction*, DOI: 10.1080/10447318.2020.1727262.

LaLomia, M. J.,& Sidowski, J. B. (1990). Measurements of computer satisfaction, literacy, and aptitudes: A review. *International Journal of Human–Computer Interaction, 2*, 231–253.

Landauer, T. K. (1997). Behavioral research methods in human–computer interaction. In M. Helander, T. K. Landauer, & P. Prabhu (Eds.), *Handbook of human–computer interaction* (2nd ed., pp. 203–227). Amsterdam: Elsevier.

Larson, R. C. (2008). Service science: At the intersection of management, social, and engineering sciences. *IBM Systems Journal, 47*(1), 41–51.

Lascu, D. & Clow, K. E. (2008). Web site interaction satisfaction: scale development consideration. *Journal of Internet Commerce, 7*(3), 359–378.

Lascu, D. & Clow, K. E. (2013). Website interaction satisfaction: A reassessment. *Interacting with Computers, 25*(4), 307–311.

Laugwitz, B., Schrepp, M. & Held, T. (2008). Construction and evaluation of a user experience questionnaire. In A. Holzinger (Ed.), *USAB 2008* (pp. 63–76). Graz, Austria, Austrian HCI and Usability Engineering Group, LNCS 5298.

Lavie, T. & Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. *International Journal of Human-Computer Studies, 60*, 269–298.

Law, E. L. & Hvannberg, E. T. (2004). Analysis of combinatorial user effect in international usability tests. In *Proceedings of CHI 2004* (pp. 9–16). ACM, Vienna, Austria.

Law, E. L., Hvannberg, E. T., Cockton, G., Palanque, P., Scapin, D., Springett, M., Stary, C., & Vanderdonckt, J. (2005). Towards the Maturation of IT Usability Evaluation (MAUSE). In M. F. Costabile & F. Paterno (Eds.), *Proceedings of INTERACT 2005* (pp. 1134–1137). IFIP, Rome, Italy.

Lewis, C. & Norman, D. (1986). Designing for error. In D. A. Norman & S. W. Draper (Eds.), *User centered system design: New perspectives on human–computer interaction* (pp. 411–432). Mahwah, NJ: Lawrence Erlbaum Associates.

Lewis, J. R. (1982). Testing small system customer set-up. In *Proceedings of the Human Factors Society 26th Annual Meeting*, Human Factors Society, Santa Monica, CA, pp. 718–720.

Lewis, J. R. (1991a). A rank-based method for the usability comparison of competing products. In *Proceedings of the Human Factors Society 35th Annual Meeting* (pp. 1312–1316). Human Factors Society, Santa Monica, CA.

Lewis, J. R. (1991b). Psychometric evaluation of an after-scenario questionnaire for computer usability studies: The ASQ. *SIGCHI Bulletin, 23*, 78–81.

Lewis, J. R. (1992). Psychometric evaluation of the post-study system usability questionnaire: The PSSUQ. In *Proceedings of the Human Factors Society 36th Annual Meeting*, Human Factors Society, Santa Monica, CA, pp. 1259–1263.

Lewis, J. R. (1993). Multipoint scales: Mean and median differences and observed significance levels. *International Journal of Human–Computer Interaction, 5*, 382–392.

Lewis, J. R. (1994). Sample sizes for usability studies: Additional considerations. *Human Factors, 36*, 368–378.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human–Computer Interaction, 7*, 57–78.

Lewis, J. R. (1996). Reaping the benefits of modern usability evaluation: The Simon Story. In G. Salvendy & A. Ozok (Eds.), *Advances in applied ergonomics: Proceedings of the 1st International Conference on Applied Ergonomics, ICAE '96* (pp. 752–757). USA Publishing, Istanbul, Turkey.

Lewis, J. R. (2001). Evaluation of procedures for adjusting problem-discovery rates estimated from small samples. *International Journal of Human–Computer Interaction, 13*, 445–479.

Lewis, J. R. (2002). Psychometric Evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human–Computer Interaction, 14*, 463–488.

Lewis, J. R. (2006). Sample sizes for usability tests: mostly math, not magic. *Interactions, 13*(6), 29–33. See corrected formula in *Interactions, 14*(1), 4.

Lewis, J. R. (2008). Usability evaluation of a speech recognition IVR. In T. Tullis & B. Albert (Eds.), *Measuring the user experience: Case studies* (pp. 244–252). Amsterdam: Morgan-Kaufmann.

Lewis, J. R. (2011a). Human factors engineering. In P. A. LaPlante (Ed.), *Encyclopedia of software engineering* (pp. 383–394). New York: Taylor & Francis.

Lewis, J. R. (2011b). *Practical speech user interface design*. Boca Raton, FL: Taylor & Francis.

Lewis, J. R. (2014). Usability: lessons learned … and yet to be learned. *International Journal of Human-Computer Interaction, 30*(9), 663–684.

Lewis, J. R. (2018a). Is the report of the death of the construct of usability an exaggeration?" *Journal of Usability Studies, 14*(1), 1–7.

Lewis, J. R. (2018b). Measuring perceived usability: The CSUQ, SUS, and UMUX. *International Journal of Human-Computer Interaction, 34*(12), 1148–1156.

Lewis, J. R. (2018c). The System Usability Scale: Past, present, and future. *International Journal of Human-Computer Interaction, 34*(7), 577–590.

Lewis, J. R. (2019a). Measuring perceived usability: SUS, UMUX, and CSUQ ratings for four everyday products. *International Journal of Human-Computer Interaction, 35*, 1404–1419.

Lewis, J. R. (2019b). Measuring user experience with 3, 5, 7, or 11 points: Does it matter? *Human Factors*, DOI: 10.1177/0018720819881312.

Lewis, J. R. (2019c). *Using the PSSUQ and CSUQ in user experience research and practice*. Denver, CO: MeasuringU Press.

Lewis, J. R., Brown, J., & Mayes, D. K. (2015). Psychometric evaluation of the EMO and the SUS in the context of a large-sample unmoderated Usability Study. *International Journal of Human-Computer Interaction, 31*, 545–553.

Lewis, J. R. & Erdinç, O. (2017) User experience rating scales with 7, 11, or 101 points: Does it matter? *Journal of Usability Studies, 12*(2), 73–91.

Lewis, J. R., Henry, S. C., & Mack, R. L. (1990). Integrated office software benchmarks: a case study. In D. Diaper et al. (Eds.), *Proceedings of the 3rd IFIP Conference on Human–Computer Interaction, INTERACT '90* (pp. 337–343). Elsevier Science, Cambridge.

Lewis, J. R. & Mayes, D. K. (2014). Development and psychometric evaluation of the Emotional Metric Outcomes (EMO) questionnaire. *International Journal of Human-Computer Interaction, 30*(9), 685–702.

Lewis, J. R. & Sauro, J. (2006). When 100% really isn't 100%: Improving the accuracy of small-sample estimates of completion rates. *Journal of Usability Studies, 3*(1), 136–150.

Lewis, J. R. & Sauro, J. (2009). The factor structure of the System Usability Scale. In M. Kurosu (Ed.), *Human centered design, HCII 2009* (pp. 94–103). Springer-Verlag, Heidelberg, Germany.

Lewis, J. R. & Sauro, J. (2017a). Can I leave this one out? The effect of dropping an item from the SUS. *Journal of Usability Studies, 13*(1), 38–46.

Lewis, J. R. & Sauro, J. (2017b). Revisiting the factor structure of the System Usability Scale. *Journal of Usability Studies, 12*(4), 183–192.

Lewis, J. R. & Sauro, J. (2018). Item benchmarks for the System Usability Scale. *Journal of Usability Studies, 13*(3), 158–167.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE—when there's no time for the SUS. In *Proceedings of CHI 2013* (pp. 2099–2102). Association for Computing Machinery, Paris, France.

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2015). Measuring perceived usability: the SUS, UMUX-LITE, and AltUsability. *International Journal of Human-Computer Interaction, 31*, 496–505.

Lilienfeld, S. O., Wood, J. M., & Garb, H. N. (2000). The scientific status of projective techniques. *Psychological Science in the Public Interest, 1*, 27–66.

Lin, H. X., Choong, Y., & Salvendy, G. (1997). A proposed index of usability: A method for comparing the relative usability of different software systems. *Behaviour & Information Technology, 16*(4/5), 267–278.

Lindgaard, G. (2006). Notions of thoroughness, efficiency, and validity: Are they valid in HCI practice? *International Journal of Industrial Ergonomics, 36*, 1069–1074.

Lindgaard, G. (2014). The usefulness of traditional usability evaluation methods. *Interactions, 21*(6), 80–82.

Lindgaard, G. & Chattratichart, J. (2007). Usability testing: What have we overlooked? In *Proceedings of CHI 2007* (pp. 1415–1424). Association for Computing Machinery, San Jose, CA.

Lucey, N. M. (1991). More than meets the i: User-satisfaction of computer systems. Unpublished thesis for diploma in applied psychology, University College Cork, Cork, Ireland.

Lusch, R. F., Vargo, S. L., & O'Brien, M. (2007). Competing through service: insights from service-dominant logic. *Journal of Retailing, 83*(1), 5–18.

Lusch, R. F., Vargo, S. L., & Wessels, G. (2008). Toward a conceptual foundation for service science: Contributions from service-dominant logic. *IBM Systems Journal, 47*(1), 5–14.

Macefield, R. (2007). Usability studies and the Hawthorne Effect. *Journal of Usability Studies, 2*(3), 145–154.

MacKenzie, I. S., & Read, J. C. (2007). Using paper mockups for evaluating soft keyboard layouts. In *Proceedings of the 2007 Conference of the Center for Advanced Studies on Collaborative Research* (pp. 98–108). Association for Computing Machinery, Richmond Hill, Canada.

MacLeod, M., Bowden, R., Bevan, N., & Curson, I. (1997). The MUSiC performance measurement method. *Behaviour & Information Technology, 16*, 279–293.

Mao, J., Vredenburg, K., Smith, P. W., & Carey, T. (2005). The state of user-centered design practice. *Communications of the ACM, 48*(3), 105–109.

Marshall, C., Brendan, M., & Prail, A. (1990). Usability of product X: Lessons from a real product. *Behaviour & Information Technology, 9*, 243–253.

Martinsa, A. I., Rosa, A. F., Queirós, A., & Silva, A. (2015). European Portuguese validation of the system usability scale. *Procedia Computer Science, 67*, 293–300.

Mayer, R. E. (1997). From novice to expert. In M. G. Helander, T. K. Landauer, & P. V. Prabhu (Eds.), *Handbook of human–computer interaction* (2nd ed., pp. 781–795). Amsterdam: Elsevier.

McCarthy, J. & Wright, P. C. (2004). *Technology as experience*. Cambridge, MA: MIT Press.

McCurdy, M., Connors, C., Pyrzak, G., Kanefsky, B., & Vera, A. (2006). Breaking the fidelity barrier: aAn examination of our current characterization of prototypes and an example of a mixed-fidelity success. In *Proceedings of CHI 2006* (pp. 1233–1242). Association for Computing Machinery, Montreal, Canada.

McDonald, S., Edwards, H. M., & Zhao, T. (2012). exploring think-alouds in usability testing: an international survey. *IEEE Transactions on Professional Communication, 55*(1), 2–19.

McDonald, S., McGarry, K., & Willis, L. M. (2013). Thinking-aloud about web navigation: The relationship between think-aloud instructions, task difficulty, and performance. In *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting* (pp. 2037–2041). HFES, Santa Monica, CA.

McDonald, S., Zhao, T., & Edwards, H. M. (2013). Dual verbal elicitation: The complementary use of concurrent and retrospective reporting within a usability test. *International Journal of Human-Computer Interaction, 29*, 647–660.

McFadden, E., Hager, D. R., Elie, C. J., & Blackwell, J. M. (2002). Remote usability evaluation: overview and case studies. *International Journal of Human–Computer Interaction, 14*, 489–502.

McGee, M. (2003). Usability magnitude estimation. In *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting* (pp. 691–695). HFES, Santa Monica, CA.

McGee, M. (2004). Master usability scaling: magnitude estimation and master scaling applied to usability measurement. In *Proceedings of CHI 2004* (pp. 335–342). ACM, Vienna, Austria,

McLellan, S., Muddimer, A., & Peres, S. C. (2012). The effect of experience on system usability scale ratings. *Journal of Usability Studies, 7*(2), 56–67.

McSweeney, R. (1992). SUMI: A psychometric approach to software evaluation. unpublished M.A. (Qual.) thesis in applied psychology, University College of Cork, Cork, Ireland.

Meuter, M. L., Ostrom, A. L., Roundtree, R., & Bitner, M. J. (2000). Self-service technologies: understanding customer satisfaction with technology-based service encounters. *Journal of Marketing, 64*, 50–64.

Michalco, J., Simonsen, J. G., & Hornbæk, K. (2015). An exploration of the relation between expectations and user experience. *International Journal of Human-Computer Interaction, 31*, 603–617.

Minge, M. & Thüring, M. (2018). The MeCUE Questionnaire (2.0): Meeting five basic requirements for lean and standardized UX assessment. In A. Marcus & W. Wang (Eds.), *Proceedings of DUXU 2018* (pp. 451–469). Springer International Publishing, Las Vegas, NV.

Minge, M., Thüring, M., Wagner, I. & Kuhr, C.V. (2016). The meCUE Questionnaire. A modular evaluation tool for measuring user experience. In M. Soares, C. Falcão, & T. Z. Ahram (Eds.), *Advances in ergonomics modeling, usability & special populations. Proceedings of the 7th Applied Human Factors and Ergonomics Society Conference 2016* (pp. 115–128). Switzerland: Springer International Press,

Moffat, B. (1990). Normalized performance ratio: A measure of the degree to which a man–machine interface accomplishes its operational objective. *International Journal of Man–Machine Studies, 32*, 21–108.

Molich, R., Bevan, N., Curson, I., Butler, S., Kindlund, E., Miller, D., & Kirakowski, J. (1998). Comparative evaluation of usability tests. In *Usability Professionals Association Annual Conference Proceedings* (pp. 189–200). Usability Professionals Association, Washington, DC.

Molich, R. & Dumas, J. S. (2008). Comparative usability evaluation (CUE-4). *Behaviour & Information Technology, 27*(3), 263–281.

Molich, R., Ede, M. R., Kaasgaard, K., & Karyukin, B. (2004). Comparative usability evaluation. *Behaviour & Information Technology, 23*, 65–74.

Molich, R., Jeffries, R., & Dumas, J. S. (2007). Making usability recommendations useful and Usable. *Journal of Usability Studies, 2*(4), 162–179.

Morse, E. L. (2000). The IUSR Project and the common industry reporting format. In *Proceedings of the ACM Conference on Universal Usability,* Association for Computing Machinery, Arlington, VA, pp. 155–156.

Nielsen, J. (1989). Usability engineering at a discount. In *Designing and using human-computer interfaces and knowledge based systems* (pp. 394–401). Amsterdam: Elsevier.

Nielsen, J. (1993). *Usability engineering*. San Diego, CA: Academic.

Nielsen, J. (1994). Usability laboratories. *Behaviour & Information Technology, 13,* 3–8.

Nielsen, J. (1997). Usability testing. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (2nd ed.). New York: Wiley.

Nielsen, J. & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the ACM INTERCHI'93 Conference* (pp. 206–213). Association for Computing Machinery, Amsterdam.

Nielsen, J. & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Conference Proceedings on Human Factors in Computing Systems, CHI '90* (pp. 249–256). Association for Computing Machinery, New York.

Nisbett, R. E. & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259.

Nørgaard, M. & Hornbæk, K. (2009). Exploring the value of usability feedback formats. *International Journal of Human-Computer Interaction, 25*(1), 49–74.

Norman, D. A. (1983). Design rules based on analyses of human error. *Communications of the ACM, 26,* 254–258.

Norman, D. A. (1986). Cognitive engineering. In D. A. Norman & S. W. Draper (Eds.), *User centered system design: New perspectives on human–computer interaction* (pp. 31–61). Mahwah, NJ: Lawrence Erlbaum Associates.

Norman, D. A. & Draper, S. W. (1986). *User centered system design: New perspectives on human-computer interaction*, Mahwah, NJ: Lawrence Erlbaum.

Nunnally, B. & Farkas, D. (2017). *UX research*, Sebastopol, CA: O'Reilly.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Ohnemus, K. R. & Biers, D. W. (1993). Retrospective versus thinking aloud in usability testing. In *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting* (pp. 1127–1131). Human Factors and Ergonomics Society, Seattle, WA.

Oliveira, T., Alhinho, M., Rita, P., & Dhillon, G. (2017). Modelling and testing consumer trust dimensions in e-commerce. *Computers in Human Behavior, 71,* 153–164.

Olmsted-Hawala, E. L., Murphy, E., Hawala, S., & Ashenfelter, K. T. (2010). Think-aloud protocols: a comparison of three think-aloud protocols for use in testing

data-dissemination web sites for usability. In *Proceedings of CHI 2010* (pp. 2381–2390). Association for Computing Machinery, Atlanta, GA.

Ostrom, A., Bitner, M., & Meuter, M. (2002). Self-service technologies. In R. Rust & P. K. Kannan (Eds.), *e-Service: New directions in theory and practice* (pp. 45–64). Armonk, NY: M. E. Sharpe.

Palmquist, R. A. & Kim, K. S. (2000). Cognitive style and on-line database search experience as predictors of web search performance. *Journal of the American Society for Information Science, 51*, 558–566.

Peres, S. C., Pham, T., & Phillips, R. (2013). Validation of the System Usability Scale (SUS): SUS in the wild. In *Proceedings of the Human Factors and Ergonomics Society 57th Annual Meeting* (pp. 192–196). HFES, Santa Monica, CA.

Pitkänen, O., Virtanen, P., & Kemppinen, J. (2008). Legal research topics in user-centric services. *IBM Systems Journal, 47*(1), 143–152.

Proctor, R. W. & Proctor, J. D. (2006). Sensation and perception. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (pp. 53–88). Hoboken, NJ: John Wiley.

Prümper, J., Zapf, D., Brodbeck, F. C., & Frese, M. (1992). Some surprising differences between novice and expert errors in computerized office work. *Behaviour & Information Technology, 11*, 319–328.

Ramli, R. & Jaafar, A. (2009). Remote usability evaluation system (e-RUE). In *Second International Conference on Computer and Electrical Engineering* (pp. 639–643). Institute of Electrical and Electronics Engineers, Dubai, UAE.

Rasmussen, J. (1986). *Information processing and human–machine interaction: An approach to cognitive engineering*. New York: Elsevier.

Rauschenberger, M., Schrepp, M., Cota, M. P., Thomashewski, J., & Olschner, S. (2013). Efficient measurement of the user experience of interactive products: How to use the User Experience Questionnaire. *International Journal of Artificial Intelligence and Interactive Multimedia, 2*(1), 39–45.

Rauschenberger, M., Schrepp, M., Olschner, S., Thomashewski, J., & Cota, M. P. (2012). Measurement of user experience: A Spanish language version of the User Experience Questionnaire (UEQ). In *Proceedings of CISTI 2012* (pp. 471–476). AISTI, Madrid, Spain.

Redish, J. (2007). Expanding usability testing to evaluate complex systems. *Journal of Usability Studies, 2*(3), 102–111.

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review, 81*, 46–54.

Reichheld, F. F. (2006). *The ultimate question: Driving good profits and true growth*, , Boston, MA: Harvard Business School Press.

Rengger, R. (1991). Indicators of usability based on performance. In H. J. Bullinger (Ed.), *Human Aspects in computing, design and use of interactive systems and*

work with terminals: Proceedings of the 4th International Conference on Human–Computer Interaction (pp. 656–660). Elsevier Science, Stuttgart, Germany.

Rohrer, C. P., Wendt, J., Sauro, J., Boyle, F., & Cole, S. (2016) .Practical Usability Rating by Experts (PURE): a pragmatic approach for scoring product usability. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, San Jose, CA, pp. 786–795.

Rosenbaum, S. (1989). Usability evaluations versus usability testing: when and why?" *IEEE Transactions on Professional Communication, 32*, 210–216.

Rowley, J. (2006). An analysis of the e-service literature: Towards a research agenda. *Internet Research, 3*, 339–359.

Rubin, J. (1994). *Handbook of usability testing: How to plan, design, and conduct effective tests*. New York: Wiley.

Rubin, J. & Chisnell, D. (2008). *Handbook of usability testing: How to plan, design, and conduct effective tests* (2nd ed.). New York: Wiley.

Rummel, B. (2014). Probability plotting: A tool for analyzing task completion times. *Journal of Usability Studies, 9*(4), 152–172.

Rummel, B. (2015). System usability scale—jetzt auch auf Deutsch. Downloaded 11/27/2017 from https://experience.sap.com/skillup/system-usability-scale-jetzt-auch-auf-deutsch/

Rummel, B. (2017). Beyond average: Weibull analysis of task completion times. *Journal of Usability Studies, 12*(2), 56–72.

Ruthford, M. A. & Ramey, J. A. (2000). Design response to usability test findings: a case study based on artifacts and interviews. In *IEEE International Professional Communication Conference* (pp. 315–323). Piscataway, NJ: IEEE Press.

Saariluoma, P. & Jokinen, J. P. P. (2014). Emotional dimensions of user experience: a user psychological analysis. *International Journal of Human-Computer Interaction, 30*, 303–320.

Sadowski, W. J. (2001). Capabilities and limitations of Wizard of Oz evaluations of speech user interfaces. In *Proceedings of HCI International 2001: Usability evaluation and interface design* (pp. 139–143). Mahwah, NJ: Lawrence Erlbaum Associates.

Safar, J. A. & Turner, C. W. (2005). Validation of a two factor structure for system trust. In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting* (pp. 497–501). Human Factors and Ergonomics Society, Santa Monica, CA.

Sauro, J. (2009). Is there a difference in usability data from remote unmoderated tests and lab-based tests? Downloaded 9/11/2019 from `https://measuringu.com/unmoderated-testing/`.

Sauro, J. (2010a). How many users do people actually test? Downloaded 2/29/2020 from `https://measuringu.com/actual-users/`.

Sauro, J. (2010b). “That's the worst website ever!”: Effects of extreme survey items. Downloaded 9/23/2010 from www.measuringusability.com/blog/extreme-items.php.

Sauro, J. (2011). *A practical guide to the system usability scale*. Denver, CO: Create Space Publishers.

Sauro, J. (2012). Predicting task completion with the system usability scale. Downloaded 2/29/2020 from https://measuringu.com/task-comp-sus/.

Sauro, J. (2013). The importance of the first choice in website navigation. Downloaded 2/29/2020 from https://measuringu.com/first-choice/.

Sauro, J. (2014). The relationship between problem frequency and problem severity in usability evaluations. *Journal of Usability Studies, 10*(1), 17–25.

Sauro, J. (2015a). *Customer analytics for dummies*. Hoboken, NJ: Wiley.

Sauro, J. (2015b). SUPR-Q: A comprehensive measure of the quality of the website User experience. *Journal of Usability Studies, 10*(2), 68–86.

Sauro, J. (2017a). Does thinking aloud affect where people look? Downloaded 2/29/2020 from https://measuringu.com/ta-gazepaths/.

Sauro, J. (2017b). Measuring usability: From the SUS to the UMUX-Lite. Downloaded 2/29/2020 from https://measuringu.com/umux-lite/.

Sauro, J. (2018a). *Benchmarking the user experience*. Denver, CO: MeasuringU Press.

Sauro, J. (2018b). Choosing the right UX testing platform. Downloaded 9/17/2019 from https://measuringu.com/ux-testing-platforms/.

Sauro J. (2018c). Do novices or experts uncover more usability issues? Downloaded 2/29/2020 from https://measuringu.com/novice-expert-issues/.

Sauro, J. (2018d). How similar are UX metrics in moderated vs. unmoderated studies? Downloaded 9/16/2019 from https://measuringu.com/moderated-vs-unmoderated/.

Sauro, J. (2018e). How to build a dedicated usability lab. Downloaded 9/10/2019 from https://measuringu.com/build-usability-lab/.

Sauro, J. (2018f). The methods UX professionals use. Downloaded 8/26/2019 from https://measuringu.com/ux-methods-2018/.

Sauro, J. (2019a). Is a three-point scale good enough? Downloaded 2/29/2020 from https://measuringu.com/three-points/.

Sauro, J. (2019b). Sample size in usability studies: How well does the math match reality? Downloaded 2/29/2020 from https://measuringu.com/sample-size-reality/.

Sauro, J. & Dumas, J. S. (2009). Comparison of three one-question, post-task usability questionnaires. In *Proceedings of CHI 2009* (pp. 1599–1608). Association for Computing Machinery, Boston, MA.

Sauro, J. & Kindlund, E. (2005). A method to standardize usability metrics into a single score. In *Proceedings of CHI 2005* (pp. 401–409). Association for Computing Machinery, Portland, OR.

Sauro, J. & Lewis, J. R. (2009). Correlations among prototypical usability metrics: evidence for the construct of usability. In *Proceedings of CHI 2009* (pp. 1609–1618). Association for Computing Machinery, Boston, MA.

Sauro, J. & Lewis, J. R. (2010). Average task times in usability tests: What to report? In *Proceedings of CHI 2010* (pp. 2347–2350). Association for Computing Machinery, Atlanta, GA.

Sauro, J. & Lewis, J. R. (2011). When designing usability questionnaires, does it hurt to be positive? In *Proceedings of CHI 2011* (pp. 2215–2223). Association for Computing Machinery, Vancouver, Canada.

Sauro, J. & Lewis, J. R. (2012). *Quantifying the user experience: Practical statistics for user research*. Waltham, MA: Morgan-Kaufmann.

Sauro, J. & Lewis, J. R. (2016). *Quantifying the user experience: Practical statistics for user research* (2nd ed.) Cambridge, MA: Morgan-Kaufmann.

Sauro, J. & Zarolia, P. (2017). SUPR-Qm: A questionnaire to measure the mobile app user experience. *Journal of Usability Studies, 13*(1), 17–37.

Schmettow, M. (2008). Heterogeneity in the usability evaluation process. In *Proceedings of the 22nd British HCI Group Annual Conference on HCI 2008: People and Computers XXII: Culture, Creativity, Interaction* (vol. 1, pp. 89–98). Association for Computing Machinery, Liverpool.

Schmettow, M. (2012). Sample size in usability studies. *Communications of the ACM, 56*(4), 64-70.

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017a). Construction of a benchmark for the User Experience Questionnaire (UEQ). *International Journal of Interactive Multimedia and Artificial Intelligence, 4*(4), 40–44.

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017b). Design and evaluation of a short version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence, 4*(4), 103–108.

Seffah, A. & Habieb-Mammar, H. (2009). Usability engineering laboratories: Limitations and challenges toward a unifying tools/practices environment. *Behaviour & Information Technology, 28*(3), 281–291.

Seffah, A., Donyaee, M., Kline, R. B., & Padda, H. K. (2006). Usability measurement and metrics: a consolidated model. *Software Quality Journal, 14*, 159–178.

Shackel, B. (1990). Human factors and usability. In J. Preece & L. Keller (Eds.), *Human–computer interaction: Selected readings* (pp. 27–41). Hemel Hempstead: Prentice Hall International.

Shneiderman, B. (1987). *Designing the user interface: Strategies for effective human–computer interaction*. Reading, MA: Addison-Wesley.

Smith, B., Caputi, P., Crittenden, N., Jayasuriya, R., & Rawstorne, P. (1999). A review of the construct of computer experience. *Computers in Human Behavior, 15,* 227–242.

Smith, D. C., Irby, C., Kimball, R., Verplank, B., & Harlem, E. (1982). Designing the Star user interface. *Byte, 7*(4), 242–282.

Smith, S. L. & Mosier, J. N. (1986). *Guidelines for designing user interface software* (Tech. Report ESD-TR-86-278) Bedford, MA: MITRE Corporation.

Sonderegger, A. & Sauer, J. (2010). The influence of design aesthetics in usability testing: effects on user performance and perceived usability. *Applied Ergonomics, 41,* 403–410.

Sonderegger, A., Schmutz, S., & Sauer, J. (2015). The influence of age in usability testing. *Applied Ergonomics, 52,* 291–300.

Soukoreff, R. W. & MacKenzie, I. S. (1995). Theoretical upper and lower bounds on typing speed using a stylus and soft keyboard. *Behaviour and Information Technology, 14,* 370–379.

Spencer, R. (2000). The streamlined cognitive walkthrough method, working around social constraints encountered in a software development company. In *Proceedings of CHI 2000* (pp. 353–359). ACM, The Hague, Amsterdam.

Spohrer, J. & Maglio, P. P. (2008). The emergence of service science: toward systematic service innovations to accelerate co-creation of value. *Production and Operations Management, 17*(3), 238–246.

Stewart, T. J. & Frye, A. W. (2004). Investigating the use of negatively-phrased survey items in medical education settings: common wisdom or common mistake? *Academic Medicine, 79*(10 Supplement), S1–S3.

Still, B. & Crane, K. (2017). *Fundamentals of user-centered design: A practical approach*. Boca Raton, FL:Taylor & Francis.

Swamy, S. (2007). How should you frame questions to measure user attitudes accurately? In N. Aykin (Ed.), *Usability and internationalization, Part II, HCII 2007* (pp. 496–505). Heidelberg: Springer.

Sy, D. (2007). Adapting usability investigations for agile user-centered design. *Journal of Usability Studies,* 2(3), 112–132.

Tedesco, D. P. & Tullis, T. S. (2006). A comparison of methods for eliciting post-task subjective ratings in usability testing. Paper presented at the Usability Professionals Association Annual Conference, UPA, Broomfield, CO.

Theofanos, M. & Quesenbery, W. (2005). Towards the design of effective formative test reports. *Journal of Usability Studies, 1*(1), 27–45.

Thimbleby, H. (2007). User-centered methods are insufficient for safety critical systems. In A. Holzinger (Ed.), *Proceedings of USAB 2007* (pp. 1–20). Heidelberg: Springer-Verlag.

Thüring, M. & Mahlke, S. (2007). Usability, aesthetics and emotions in human-technology interaction. *International Journal of Psychology*, *42*(4), 253–264.

Tohidi, M., Buxton, W., Baecker, R., & Sellen, A. (2006). getting the right design and the design right: testing many is better than one. In *Proceedings of CHI 2006* (pp. 1243–1252). Association for Computing Machinery, Montreal, Canada.

Tractinsky, N. (2017). The usability construct: A dead end? *Human-Computer Interaction, 33*(2), 131–177.

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers, 13*, 127–145.

Tuch, A. N., Roth, S. P., Hornbæk, K., Opwis, K., & Bargas-Avila, J. A. (2012). Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in HCI. *Computers in Human Behavior*, *28*(5), 1596–1607.

Tullis, T. & Albert, W. (2013). *Measuring the user experience: Collecting, analyzing, and presenting usability metrics* (2nd ed.). Waltham, MA: Morgan-Kaufmann.

Tullis, T., Fleischman, S., McNulty, M., Cianchette, C., & Bergel, M. (2002). An empirical comparison of lab and remote usability testing of web sites. In *Proceedings of the Usability Professionals Association* (pp. 1–8). Usability Professionals Association, Chicago, IL.

Tullis, T. S. & Stetson, J. N. (2004). A comparison of questionnaires for assessing website usability. Paper presented at the UPA Annual Conference. home.comcast.net/~tomtullis/publications/UPA2004TullisStetson.pdf.

Turner, C. W., Lewis, J. R., & Nielsen, J. (2006). Determining usability test sample size. In W. Karwowski (Ed.), *The international encyclopedia of ergonomics and human factors* (pp. 3084–3088). Boca Raton, FL: CRC Press.

Uldall-Espersen, T., Frøkjær, E., & Hornbæk, K. (2008). Tracing impact in a usability improvement process. *Interacting with Computers*, *20*, 48–63.

van de Vijver, F. J. R., & Leung, K. (2001). Personality in cultural context: Methodological issues. *Journal of Personality, 69*, 1007–1031.

van den Haak, M. J., & de Jong, D. T. M. (2003). Exploring two methods of usability testing: Concurrent versus retrospective think-aloud protocols. In *Proceedings of the International Professional Communication Conference, IPCC 2003* (pp. 285–287). Institute of Electrical and Electronics Engineers, Orlando, FL.

van den Haak, M. J., & de Jong, M. D. T. (2005). Analyzing the interaction between facilitator and participants in two variants of the think-aloud method. In *2005*

IEEE International Professional Communication Conference Proceedings (pp. 323–327). Institute of Electrical and Electronics Engineers, Limerick, Ireland.

van den Haak, M. J., de Jong, M. D. T., & Schellens, P. J. (2006). Constructive interaction: an analysis of verbal interaction in a usability setting. *IEEE Transactions on Professional Communication, 49*(4), 311–324.

Venkatesh, V. (2000). Determinants of perceived ease of use: integrating control, intrinsic motivation, and emotion into the technology acceptance model. *Information Systems Research, 11*(4), 342–365.

Venkatesh, V. & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences, 39*(2), 273–315.

Vilbergsdóttir, S. G., Hvannberg, E. T., & Law, E. L. (2006). Classification of usability problems (CUP) scheme: Augmentation and exploitation. In *Proceedings of NordiCHI 2006* (pp. 281–290). Association for Computing Machinery, Oslo, Norway.

Virzi, R. A. (1990). Streamlining the design process: running fewer subjects. In *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 291–294). Human Factors Society, Santa Monica, CA.

Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors, 34*, 457–468.

Virzi, R. A. (1997). Usability inspection methods. In M. G. Helander, T. K. Landauer, & P. V. Prabhu (Eds.), *Handbook of human–computer interaction* (2nd ed., pp. 705–715). Amsterdam: Elsevier.

Virzi, R. A., Sorce, J. F., & Herbert, L. B. (1993). A comparison of three usability evaluation methods: Heuristic, think-aloud, and performance testing. In *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting* (pp. 309–313). Human Factors and Ergonomics Society, Santa Monica, CA.

Vredenburg, K. (2003). Building ease of use into the IBM user experience. *IBM Systems Journal, 42*(4), 517–531.

Vredenburg, K., Isensee, S., & Righi, C. (2002). *User-centered design: An integrated approach*. Upper Saddle River, NJ: Prentice Hall.

Vredenburg, K., Mao, J. Y., Smith, P. W., & Carey, T. (2002). A survey of user centered design practice. In *Proceedings of CHI 2002* (pp. 471–478). Association for Computing Machinery, Minneapolis, MN.

Watson, D., Clark, A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS Scales. *Journal of Personality and Social Psychology, 54*(6), 1063–1070.

Wenger, M. J. & Spyridakis, J. H. (1989). The relevance of reliability and validity to usability testing. *IEEE Transactions on Professional Communication, 32*, 265–271.

West, R. & Lehman, K. R. (2006). Automated summative usability studies: An empirical evaluation. In *Proceedings of CHI 2006* (pp. 631–639). Association for Computing Machinery, Montreal, Canada.

Whiteside, J., Bennett, J., & Holtzblatt, K. (1988). Usability engineering: Our experience and evolution. In M. Helander (Ed.), *Handbook of human–computer interaction* (pp. 791–817). Amsterdam: North-Holland.

Wickens, C. D. (1998). Commonsense statistics. *Ergonomics in Design, 6*(4), 18–22.

Wiethoff, M., Arnold, A., & Houwing, E. (1992). *Measures of cognitive workload,* MUSiC ESPRIT Project 5429 document code TUD/M3/TD/2.

Wildman, D. (1995). Getting the most from paired-user testing. *Interactions, 2*(3), 21–27.

Williams, G. (1983). The Lisa computer system. *Byte, 8*(2), 33–50.

Winter, S., Wagner, S., & Deissenboeck, F. (2008). A comprehensive model of usability. In *Engineering Interactive Systems* (pp. 106–122). International Federation for Information Processing, Heidelberg.

Wixon, D. (2003). Evaluating usability methods: Why the current literature fails the practitioner. *Interactions, 10*(4), 28–34.

Woolrych, A., & Cockton, G. (2001). Why and when five test users aren't enough. In J. Vanderdonckt, A. Blandford, & A. Derycke (Eds.), *Proceedings of IHM–HCI 2001 Conference* (vol. 2, pp. 105–108). Toulouse, France: Cépadèus Éditions.

Wright, R. B. & Converse, S. A. (1992). Method bias and concurrent verbal protocol in software usability testing. In *Proceedings of the Human Factors and Ergonomics Society 36th Annual Meeting* (pp. 1220–1224). Human Factors and Ergonomics Society, Santa Monica, CA.

Wu, J., Chen, Y., & Lin, L. (2007). Empirical evaluation of the revised end user computing acceptance model. *Computing in Human Behavior, 23*, 162–174.

Xue, M., & Harker, P. T. (2002). Customer efficiency: Concept and its impact on e-business management. *Journal of Service Research, 4*(4), 253–267.

Yusop, N. S. M., Grundy, J., & Vasa, R. (2017). Reporting usability defects: A Systematic literature review. *IEEE Transactions on Software Engineering, 43*(9), 848–867.

Zijlstra, R., & van Doorn, L. (1985). *The construction of a scale to measure subjective effort* (Tech. Report). Department of Philosophy and Social Sciences, Delft