Usability Data Quality

David R. Danielson
Stanford University, USA

INTRODUCTION

A substantial portion of usability work involves the coordinated collection of data by a team of specialists with varied backgrounds, employing multiple collection methods, and observing users with a wide range of skills, work contexts, goals, and responsibilities. The desired result is an improved system design, and the means to that end are the successful detection of, and reaction to, real deficiencies in system usability that severely impact the quality of experience for a range of users.

In the context of user-centered design processes, valid and reliable data from a representative user sample is simply not enough. High-quality usability data is not just representative of reality. It is useful. It is persuasive in the eyes of the right stakeholders. It results in verifiable improvements to the system for which it is intended to represent a deficiency. The data must be efficiently and effectively translated into development action items with appropriate priority levels, and it must result in effective work products downstream, leading to cost-effective design changes.

The remainder of this article (a) briefly reviews basic usability data collection concepts, (b) examines the dimensions that make up high-quality usability data, and (c) suggests future trends in usability data quality research.

BACKGROUND

Usability data are critical to the successful design of systems intended for human use, and are defined by Hilbert and Redmiles (2000) as any information used to measure or identify factors affecting the usability of a system being evaluated. Such data are collected via *usability evaluation methods* (UEMs), methods or techniques that can assign values to usability dimensions (J. Karat, 1997) and/or indicate usability deficiencies in a system (Hartson, Andre, & Williges, 2003). Usability dimensions are commonly taken to include at least user efficiency, effectiveness, and subjective satisfaction with a system in performing a specified task in a specified context (ISO 9241-11, 1998), and frequently also include system memorability and learnability (Nielsen, 1993a).

Usability data are collected using either *analytic methods*, in which the system is evaluated based on its interface design attributes (typically by a usability expert), or *empirical methods*, in which the system is evaluated based on observed performance in actual use (Hix & Hartson, 1993). In *formative evaluation*, data are collected during the development of a system in order to guide iterative design. In *summative evaluation*, data are collected to evaluate a completed system in use (Scriven, 1967). Usability data have been classified in numerous other models and frameworks frequently focusing on the procedure for producing the data (including the resources expended and the level of the formality of the method), the (relative) physical location of the people and artifacts involved, the nature and fidelity of the artifact being evaluated, and the goal of the collection process.

DIMENSIONS OF USABILITY DATA QUALITY

Usability-data quality refers to the extent to which usability data efficiently and effectively (a) predict system usability in actual usage (validity, reliability, representativeness, and completeness), (b) can be analyzed (communicative effectiveness and efficiency, and analyst estimates of severity), and (c) can be reacted to (downstream utility, impact, and cost effectiveness). This section discusses the dimensions of usability data quality and their assessment.

Validity

High-quality usability data are predictive of a real deficiency in one or more usability attributes for a given system. End-user behavior and comments may be perfectly unbiased or unaffected by the collection process, yet still lack validity from the perspective of usability science. Strict performance measures (such as time on task) may be viewed as lacking validity primarily because they often fail to, on their own, demonstrate an underlying problem (Gediga, Hamborg, & Düntsch, 2002). Qualitative data more often do point directly to a deficiency, but if a user comments on a system feature that will never be used, for example, the comment may truly reflect the user's attitudes but nonetheless lack validity.

Verifying usability data validity lies in comparing the data's predicted problems to the actual system performance in use (John & Marks, 1997; Nielsen & Phillips, 1993). In practice, assessing validity is nontrivial for three fundamental reasons. First, there is not widespread agreement on how to operationalize ultimate usability criteria into actual criteria (Gray & Salzman, 1998; Hartson et al., 2003); that is, agreeing on standard measures (and measurement procedures) for the underlying dimensions of usability itself is a long-standing difficulty. Second, observing the system in use and recording deficiencies is itself a usability data collection process, and thus at best the actual criterion is subject to possible validity concerns of its own. While these first two problems are by no means unique to usability research, they illustrate the difficulty in assessing usability data quality without a widely agreed upon method for identifying what will be accepted as the system's real deficiencies. Finally, individual pieces of usability data are often difficult to translate into underlying problems, and this step is necessary if validity is to be assessed.

To make the problem slightly more tractable, researchers have by and large elected to evaluate validity using usability testing as a benchmark for comparison, as it is assumed to most closely reflect system performance in use (Cuomo & Bowen, 1994; Desurvire, 1994; Jacobsen, Hertzum, & John, 1998). There are of course potential problems with this approach as usability testing has at least ecological validity concerns (Thomas & Kellogg, 1989). Indeed, this problem generally makes the literature comparing UEM effectiveness difficult to interpret (Gediga et al., 2002; Gray & Salzman, 1998). Ideally, a standard method is applied to assessing live system performance, producing a usability problem set. Validity is then assessed by comparing the problem set produced by a UEM to the standard set (Sears, 1997).

Reliability and Representativeness

High-quality usability data not only indicate real problems, but indicate problems that will be repeatedly encountered by individual users (reliable) and by a wide range of users (representative). As with many disciplines, data collected for usability purposes vary in the extent to which the repeated exposure to a problem is a good predictor of validity. While subjective satisfaction ratings that vary one day to the next put validity in question, encountering only occasional difficulty in executing a system action or completing a task, for example, does not since user errors indicating real interface problems commonly vary in frequency of occurrence. Unlike research in many other disciplines, representativeness across participants is not simply a question to be investigated, but a contributor to problem importance and therefore a dimension of data quality.

Measuring reliability and representativeness is a matter of identifying the recurrence of specific problems (Jeffries, Miller, Wharton, & Uyeda, 1991). Such measurement is nontrivial because problem reports may differ in verbiage but still indicate the same underlying problem, or conversely may be similar in their qualitative descriptions but indicate different deficiencies (Andre, Hartson, Belz, & McCreary, 2001; Hartson et al., 2003).

Completeness

High-quality usability data represent usability problems in their entirety. One of the critical difficulties in analyzing pure behavioral data is their lack of contextual information about the user's current task, attention level, and cognitive processes while a problem takes place (Hilbert & Redmiles, 1999); another problem is their flood of extraneous data that are not useful in evaluating the deficiency (Hartson & Castillo, 1998). Ideal usability data predict a