

# How Many Users Are *Really* Enough...And More Importantly *When*?

Michael A. Katz & Christian Rohrer

Yahoo! Inc  
701 First Ave.  
Sunnyvale, CA 94089 USA

## ABSTRACT

While some practitioners have argued that five users are enough to conduct a usability study, others advocate larger sample sizes or formulas to determine the appropriate number. Although productive, this debate has largely ignored the distinction between formative and summative research leaving many practitioners unable to clearly articulate the circumstances that determine whether a small or large sample is required. This has led to an overemphasis of quantitative measures at the expense of qualitative insight and the specific practice of relying on numerous observations of a usability issue to establish validity. In our view, accounts of user difficulty that include a description of the problem along with its potential cause and impact do not require large sample sizes to drive meaningful design change. By addressing arguments central to this debate, we intend to clarify the appropriate uses of the usability study methodology and improve the credibility and impact of usability professionals in practical settings.

## Author Keywords

Usability, number of participants, formative research, summative research

## ACM Classification Keywords

H5.2. Information interfaces and presentation (e.g., HCI): User Interfaces.

## INTRODUCTION

A spirited debate has emerged in recent years concerning the number of participants required to adequately test the usability of a system [26, 23, 18, 7]. While some have argued that five users is enough [18, 24], others have advocated larger sample sizes and formulas to determine

the appropriate number of participants for a study [23, 26]. What all such accounts have in common is the implicit assumption that a certain proportion of existing usability issues must be discovered to make a usability study worthwhile:

“A cost/benefit balance must be used to determine how many users should test a system. If more than necessary are used, the cost of extra users will outweigh the benefits of the knowledge gained. Conversely, too few test users may miss key problems that render a system close to unusable. *A magic formula is needed to tell us that x users are needed to find y% of problems.*” [26, p. 105, emphasis added]

While we consider this debate worthwhile, we feel that its applicability has been generalized inappropriately to all usability studies (regardless of their intended purpose) and has in the process clouded the value of the usability study methodology. Given the goal of the “How many users is enough?” debate, we consider this turn of events to be ironic in that it has shifted the focus of usability studies away from their primary goal, namely to improve the quality of products [25].

When debating the question of “How many users is enough?” we feel that it is important to clarify the distinction between usability studies intended to assess products (i.e., summative research) and those intended to improve them (i.e., formative research) [19]. It is also necessary to understand the relationship between the number of participants required to *discover* all the existing usability issues with a product and the number of participants required to *validate* the existence of a specific usability issue. To date, these distinctions have not been made clear [16] which has led to the gross misconception of usability studies as requiring a minimum number of participants to be worthwhile, and has diminished the perceived value of qualitative insight provided by the usability study methodology.

Given the prominence of the “How many users is enough?” debate, we decided to take this opportunity to address some popular misconceptions of usability studies that have the

aim of *improving* products, misconceptions that have eroded the practical impact of usability research in recent years. By focusing attention on the key benefits of usability research, the mischaracterized validity of usability issues, and the critical role of the usability practitioner in reporting findings, we hope to guide usability professionals toward improved influence and impact in practical settings.

### **ASSESSING THE QUALITY OF A PRODUCT: DISCOVERING THE UNIVERSE OF USABILITY PROBLEMS**

To understand the problem with the debate regarding sample size of usability studies, one must consider the different ways in which usability studies are utilized, namely assessment and improvement of products. By “assessment,” we refer to the goal of providing a metric of the quality of a product that can be used to benchmark the product against later versions or competitors (i.e., summative research). By “improvement,” we refer to the goal of revealing and addressing usability problems discovered with the product and reducing the risk of failure when the product is introduced into the marketplace (i.e., formative research). To date, this distinction has not been made clear among practitioners:

“...an evaluator may choose to run a small number of subjects *only if he or she is convinced that there is not an inordinate amount of risk that the evaluation will fail to detect major usability problems...*if some severe usability problems are likely to be missed when the sample size is small, an evaluator may be less likely to restrict the sample size for fear that a critical problem will evade detection.” [24, p. 460, emphasis added]

Determining how many participants in a study is “enough” is a question that is appropriate when the goal of the usability study is one of assessment. To assess the quality of a product, one must have confidence that all of the major usability problems are known and a sufficient number of participants must therefore be tested to satisfy this requirement [17].

As an example, consider the following thought experiment. Imagine a new product with three distinct major flaws that must be discovered (and corrected) during a usability study for the product to be successful in the marketplace. Assume hypothetically that two of the major flaws would be revealed by participants 1 and 3 respectively, while the final major flaw would be revealed by participant 15. If the goal of the usability study is to assess the quality of the product, then fifteen participants would be the minimum required to confidently assert that the product would be successful. Given this goal of *assessment*, testing fourteen participants would be insufficient as it would lead to an inaccurate assessment of the quality of the product.

### **IMPROVING THE QUALITY OF A PRODUCT: RECONSIDERING THE SCENARIO**

However, the primary mission of usability professionals in practical settings is often not simply to determine *if the product will fail* upon introduction to the marketplace but to *reduce the risk of failure*. When the goal of a study is one of improvement, the requirement of discovering all major usability problems no longer applies.

Although this assumption has been challenged by some, many practitioners have come to the defense of the tradeoff between discovering all usability issues with a design and being able to conduct more iterations in the design process [25, 1]. While testing a small number of users may not be enough to accurately assess the quality of a product, it provides an effective practical basis for improving products through iterative design and testing. Furthermore, if a usability investigation focuses on the most important target users and the most critical tasks, then it is likely that usability problems revealed will be important to the success of the product regardless of whether *all* problems are discovered [13].

In our view, revealing and addressing usability problems with a product will yield an improved product, and whether or not *all* of the major usability issues have been discovered is irrelevant to the claim that the product has been improved [15]. As any usability issue may translate to fewer customers, lower levels of satisfaction, or a damaged brand, every usability issue addressed will lead to reduced risk of failure to the product.

Consider the above scenario once more, except with the new goal of improving the product relative to its current state. If only three participants are tested, then two of the three major flaws will be discovered (and presumably addressed). If two-thirds of the major flaws with the product were addressed, would it not be fair to assert that the risk of failure of the product were reduced and the study were worthwhile?

### **SAMPLE SIZE AS A MEANS OF ENSURING VALIDITY OF THE FINDINGS**

A related issue is the common belief that a minimum number of participants would be required (for any usability study) to ensure a degree of validity in the findings. For example, an issue revealed by participant number 3 in a study may not be considered an issue by the time you reach participant 15. Therefore, if less than fifteen participants are tested, the findings would be in question.

We consider this line of reasoning to be misguided and reflective of a misunderstanding of appropriate criteria for determining the validity of usability issues. As we will argue in detail below, the validity of a usability issue depends not on the number of participants who exhibited the issue, but rather the ability of the usability professional to create a plausible and rational account of the exhibited behavior.

As others have pointed out, the number of participants required to establish validity for research findings depends to a great extent on the goals of the research [5]. When the goal is to gather qualitative insights to improve products, large sample sizes are not required:

“When doing any kind of user research you can study large numbers shallowly or small numbers in depth (*which method you need depends upon your questions and the kinds of answers you are looking for*).” [5, p. 32, emphasis added]

### **FROM PARTICIPANT DATA TO USABILITY ISSUES: THE COMMON PRACTICE OF FOCUSING ON TRENDS AND PATTERNS**

Over the course of a usability study, the researcher observes many behaviors and listens to many participant comments. A difficult aspect of being a usability professional is then determining which behaviors and comments are worthy of reporting to stakeholders as being reflective of usability “issues” with the product. As a usability professional, she is aware that she must be cautious about what to report to stakeholders as her credibility may be challenged.

Common practice at this point is for the usability professional to focus on patterns or trends in user behavior, and texts describing basic usability techniques often reference the importance of high-frequency behaviors or common areas of confusion among participants as a means of calling out key usability problems [10, 21, 14]:

“Look for repetition and things that may be caused by common underlying problems... Having grouped all the observations, go through the groups and consolidate them, separating the groups of unrelated topics. *Throw away those that only have one or two individual observations.*” [10, p. 296, emphasis added]

Following this line of reasoning, if nearly all users exhibited difficulty then surely a usability issue must exist, but if difficulty were experienced by only a few users then it is inappropriate to conclude that a usability issue would exist in the entire user population. In the latter case, claiming a usability issue would encourage fierce debate among stakeholders and may potentially damage the credibility of the usability professional, leading the issue to be left unreported.

Despite the prevalence of this practice in the usability community, we feel that it represents a fundamental misunderstanding of the usability study methodology and such instances (in our view) have eroded the impact of usability professionals in corporate or other practical settings.

The above scenario assumes incorrectly that a usability study is *intended* to reflect the behavior of the entire population of users. Given that framework, only behaviors exhibited by a large proportion of the sample are considered

to be “issues” in that they are reflective of the population. However, the goal of usability studies is not to represent the population, but rather to expose potential areas of confusion or difficulty when using a product and to address those areas. Given this perspective, the *number* of participants who exhibited difficulty is of relatively little significance.

If one accepts our claim that frequency of an error (or area of confusion) is not required to validate an issue, then what makes for a usability “issue?” In our view, an interpretation of participant behavior is worthy of reporting to stakeholders when the following conditions are met:

- An instance of confusion occurs that can be described (in terms of the behavior that occurred).
- A reasonable argument can be made regarding the cause of the difficulty.
- The impact of the difficulty (on user success or satisfaction) can be clearly described.

What is notably missing from the above set of criteria is any mention of frequency of occurrence. Instead, the focus is on *telling a story* about participant behavior. (These arguments are not new; others have advocated focusing on the qualitative nature of usability issues and their impact rather than the frequency of their occurrence) [2, 15, 25, 20, 19, 4, 22].

### **“TELLING A STORY”: HOW PROVIDING A GOOD ACCOUNT OF USER BEHAVIOR CAN MAKE SAMPLE SIZE IRRELEVANT**

In their discussion of Rapid Iterative Testing and Evaluation (RITE), Medlock et al. [15] argued that in certain cases *one* participant can be enough to reveal a usability issue worthy of reporting to stakeholders. As an example, they presented the hypothetical case in which one participant in a usability study failed to notice the difference between a red and green color code. If it is known that the participant is red-green color blind, then additional participants are not needed to demonstrate that the red-green color states present a problem [20]:

“With respect to reliability/validity, making a change based on minimal participants is a risk but one we are willing to take in certain situations. *When the problem and solution are “obvious”...seeing an issue once is sufficient.*” [15, emphasis added]

We agree with this viewpoint but we also feel that usability findings based on the behavior of one participant are not as extreme as implied by Medlock et al. While we agree that there is risk to the approach of advocating findings based on the behavior of one participant, we believe that this risk is no greater than that associated with defending any usability finding to an audience that understands the qualitative nature of usability research. Unfortunately, however, this level of understanding is uncommon among product stakeholders (and many researchers as well).

## Personals Mailbox

[Mailbox Home](#) | [Icebreaker Home](#) | [Drafts](#) | [Trash](#) | [Options](#) | [Help](#)

Response to my profile: Looking to make a connection

[Reply](#) [Delete](#) [Block User](#)

[Previous](#) | [Next](#) | [View Reply Box](#)

Date: 05/16/2004  
From: user\_one  
To: utest\_f  
Subject: Hi!

Hi. My name is Bill and I'm 44 years old. I thought your profile sounded really interesting. I'm divorced with no kids but I'd like to have some in the future. Anyway, drop me a note if you get a chance.



Looking to start a family..  
( Active during the last 24 hours! )

Personals Member  
Age: 44; Los Angeles, CA  
I'm divorced with no kids but I'm looking for someone who wants to start a family with someone special. I hope to... [More](#)

[Reply](#) [Delete](#) [Block User](#)

[Previous](#) | [Next](#) | [View Reply Box](#)

Figure 1. Message page.

To some in the usability profession, testing with a small number of users at low cost inevitably leads to poor data and interpretation [3]. However, this argument is weakened by the criteria presented above for classifying usability issues along with the assumption that reported usability issues would be presented to stakeholders as complete accounts of user behavior (including cause and impact). In our view, such criticisms of discount usability are valid only when usability professionals rely on nothing more than their opinion or prior experience to classify usability issues. (On this point, research exists that strongly suggests that all usability professionals are not equally skilled.) [16]

Furthermore, we disagree with attacks on the validity of discount usability research that are based on the absence of sufficient “experimental controls” to establish causation [3]. As inferential statistics are not being used to determine cause and effect, discount usability methods do not require experimental control in the classic sense to be valuable. As many usability professionals will attest, much can be learned by watching users interact with an interface, including an understanding of factors that may have led to their behavior and possible solutions to remedy errors.

In fact, it is our position that in cases where a behavior can be clearly described along with a plausible account for its cause and impact, then the sample size for that finding is irrelevant as pertains to the validity of the finding.<sup>1</sup> (The

---

<sup>1</sup> These arguments do not apply to user preference data which require larger sample sizes to be of any practical significance. Unlike usability data, preference data is quantitative and *is* intended to represent the entire population of users. However, preference data obtained from small samples of users can be useful in providing qualitative insight into the motivations and drivers of user preferences. (E.g., “Design A was preferred for its breadth of menu options and unobtrusive advertising.”)

usability professional can provide what statistics cannot, namely an account of user behavior that provides insight into how a product may be improved.)

The basic elements of such an account are presented below, followed by an example:

- A participant representative of the target users for the product.
- A difficulty that stemmed from a behavior that was reasonable given the product domain.
- A clear description of the problem or difficulty.
- A clear description of the impact of such difficulty.
- A rational account of the cause of the problem.
- Suggestions as to how the product may be improved based on the account.

For the example provided below<sup>2</sup>, ask yourself how you would perceive the usability issue differently depending on the size of the sample that exhibited it.

### TELLING A STORY FOR A USABILITY ISSUE: EXAMPLE

In 2003, a usability study was conducted for Yahoo! Personals to investigate the process of communicating with a user who responds to a posted ad. The intended flow was simple: A user would receive a message from a potential suitor and if interested would reply to that message by clicking the “Reply” button (see Figure 1 above). Upon doing so, the user would be presented with the ability to reply to the sender at no cost.

---

<sup>2</sup> The personal information contained in the example was altered so as to protect the privacy of the individual portrayed in the ad. Thanks to Jeralyn Reese for providing the example.

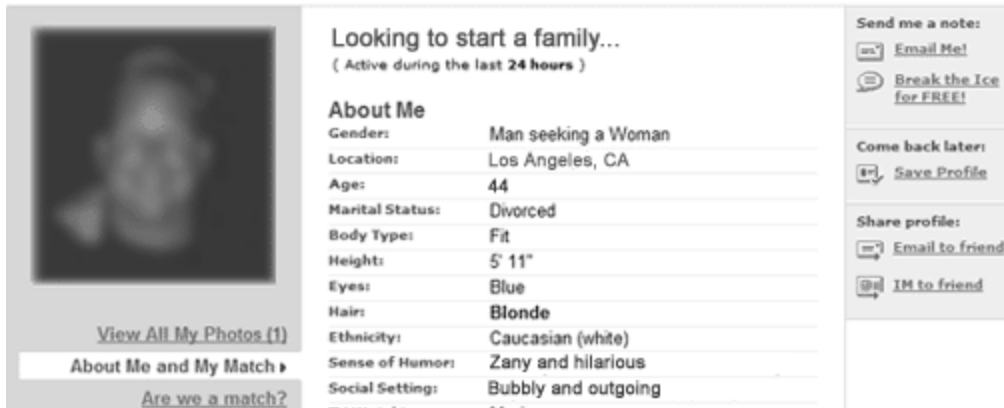


Figure 2. Ad Detail page.

However, instead of clicking the “Reply” button, the first participant in the study clicked the “More” link in the member description box on the right to learn more about the sender. This led to the Ad detail page (see Figure 2 above).

At this point, the participant clicked the “Email Me!” link in the rightmost section to reply to this sender’s original message and was presented with the Subscription page (see Figure 3 below) which led her to incorrectly conclude that a paid subscription to Yahoo! Personals were required to reply to the sender. Given this participant behavior, it was then up to the researcher to determine if there existed a genuine usability issue worthy of reporting to product stakeholders. To do this, the researcher sought to identify the nature of the usability problem, its impact, its cause, and potential solutions. If a plausible and rational account of the issue could be offered, then it would be worthy of reporting to stakeholders.

**TELLING A STORY FOR A USABILITY ISSUE: YAHOO! PERSONALS**

Presented below are the questions considered by the researcher (listed earlier) and the answers as pertained to this issue.

**Q.** Was this participant a **representative target user** for this service?

**A.** Yes. This participant met the demographic, behavioral, and attitudinal criteria for participation in the study.

**Q.** Did the participant **behave in a reasonable way** given this domain?

**A.** Yes. The actions the participant took to accomplish the goal of replying to an ad were reasonable and not unusually deviant in any way. For example, the participant did not attempt to navigate to Yahoo! Maps or demonstrate a misunderstanding of the goal to be accomplished.

**Q.** What was the **nature of the problem**?

**A.** The participant assumed incorrectly that a paid subscription to Yahoo! Personals were required to reply to a sender’s message.

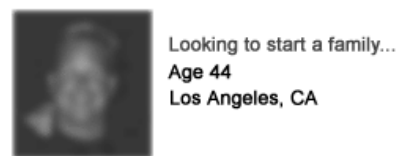
**Q.** What was the **impact of the problem**?

**A.** The participant failed to reply to the sender’s message, a task defined by product stakeholders as critical to success of the product.

**Q.** What was the **cause of the problem**?

**A.** Yahoo! Personals was designed to cater to two different use case scenarios: (1) Replying to a sender’s message and (2) initiating communication with a person based on search results. The intended flow of the first case was through use of the “Reply” button which allows a user to respond at no cost. However, the intended flow of the second case was for users to click “Email Me!” from an Ad detail page found via a search on the Personals site, at which point a user would be prompted to subscribe to the service.

**Want to contact this person?**



Subscribe now to email or instant message as many singles as you wish!

- 1 month for \$19.95
- 3 months for \$42.95 - less than \$15/month!
- 12 months for \$89.95 - less than \$8/month!

The subscription plan you choose will **automatically renew** using your credit card until you decide to cancel. Please see additional terms on the next page.

**Start Now!**

Figure 3. Subscription page.

The participant described above intended to accomplish case 1 but took the flow intended for case 2. This error was made possible through the presence of the “Email Me!” link on the Ad detail page (Figure 2 above) which led users to the Subscription page (Figure 3 above) implying an associated fee.

**Q.** How might this problem be **addressed**?

**A.** As users may wish to learn more about a sender prior to contacting them, the information provided by the Ad detail page should be attainable without disrupting the flow of use case scenario 1. While the “Email Me!” link (on the Ad detail page) should be available to users initiating communication based on search results (use case scenario 2), it should not be available for users intending to reply to a sender’s message (use case scenario 1). In other words, the system should accurately reflect the specific context of use.<sup>3</sup>

The key point of this example is that the sample size for the finding is irrelevant to the validity of the finding. Was the issue at hand unclear? Was the cause not a rational one? While the number of participants who committed the behavior can be useful to determine how the issue should be prioritized among all issues, it has no bearing on the validity of the issue itself. (Would it really be necessary to observe similar behavior in other participants before classifying this behavior as a usability issue?)

Having provided plausible and rational answers to all of the above questions, the researcher was then left to convince the product stakeholders that a usability issue existed that should be addressed. This notion of “selling the story” is described below.

#### **FROM TELLING THE STORY TO SELLING THE STORY**

As with any usability finding, the story must not only be told, but *sold*. The best a researcher can do is to present the nature of the finding along with a plausible account of its cause and impact and hope that the product stakeholders will agree.<sup>4</sup> Relying on the sample size to defend the finding (whether it be “eight” or “more than half”) is tantamount to relying on the basic misunderstandings of usability research possessed by many product professionals. Observing many participants committing the same error can be powerful data, but this data should be used to solidify the account of the behavior rather than provide justification for

---

<sup>3</sup> The billing model for Yahoo! Personals has changed since the completion of this usability study.

<sup>4</sup> Fortunately, there are tools at the disposal of researchers to help sell the story for a finding. Perhaps the most powerful among such tools is the video clip [9]. Presenting a video playback of a participant experiencing difficulty due to specific design factors can have a powerful effect on stakeholders. Often in such situations, the issue of sample size is not even mentioned.

the behavior as a “usability issue.” *The only circumstance for which a sample of one is “not enough” is when the behavior of the one participant does not make clear the cause and/or impact of the usability issue.*

Therefore, instead of reluctantly advocating findings based on the behavior of one participant, we embrace them as they most accurately represent the true nature of usability studies as being dependent upon qualitative accounts of user behavior rather than reflections of user populations.

#### **USING FREQUENCY DATA TO CLASSIFY AND PRIORITIZE ISSUES**

While we agree that frequency and severity data is often important when conducting usability research, we feel that its value and applicability has been distorted by many in the usability profession. The following quote is reflective of current common thinking:

“It is not enough to simply find problems – they must be understood and prioritized, *and neither is possible without good frequency and severity data.*” [26, emphasis added, p. 108]

We disagree with the statement above from Woolrych and Cockton [24, 17, 12, 8] that frequency data is *required* to understand and prioritize problems. The nature of the problem itself relative to the goals of product stakeholders along with the account of the problem provided by the researcher are in many cases sufficient to understand and prioritize the problem in question. As elucidated by the earlier example, one participant can clearly be enough to understand a problem with a product.

In contrast to the view of Woolrych and Cockton, we believe that the true value of frequency data is to ensure the quality of the account of the problem that is reported by the usability researcher. While testing several participants may be unnecessary to reveal certain usability issues, numerous participants may be required to fully *understand* a usability issue and tell a story about it (in accordance with the above criteria). If a participant in a usability study (who is a target user of the product) commits an error which results in total failure to complete a key task, then a usability issue must exist. However, it is not always the case that the cause of the difficulty will be apparent to the researcher [18].

In such cases, testing more users is helpful in that other users may exhibit the same difficulty while simultaneously revealing key insights as to the cause of the confusion. In this regard, the number of participants is relevant to the extent that it helps clarify the nature of usability issues. However, we consider it inappropriate to use frequency data from qualitative research as the primary basis for prioritizing issues.

For example, which of the following two usability issues deserves higher priority?

- A problem that resulted in moderate difficulty for 9 of 12 participants
- A problem that resulted in failure for 3 of 12 participants

This example illustrates the futility of relying on frequency data to prioritize usability issues. Given the small sample size (twelve in this example), focusing on the number of participants who experienced a problem clouds the fact that three participants experienced a severe problem with using the product that resulted in failure, a problem for which the usability professional could provide a reasonable account as to its cause.

In our view, if a participant behaving reasonably for a given domain reveals a problem that can be explained with a reasonable account (in terms of its cause and impact), then no other information is required to understand and prioritize that problem. Given the business goals of product stakeholders, this account is sufficient to classify and prioritize the problem accordingly. As an illustration, we can extend the Yahoo! Personals example provided earlier to reflect the role of business goals in determining how to prioritize the usability issue:

**Q.** How important is the issue given the business goals of product stakeholders?

**A.** The ability to reply to the sender’s message (a free service) was specified by the product stakeholders as a key business goal as it has implications for both users who post ads as well as those who respond to ads. Users who respond to ads must pay to do so and their continued use of the Yahoo! Personals service depends on the ability of interested suitors to communicate with them easily. Similarly, it is important for users to post ads so as to increase the likelihood that a potential suitor will find an ad of interest.

The usability issue in question may affect a user’s likelihood to respond to future ads as he may interpret the lack of replies as a indication of a low quality service. Similarly, it may affect the likelihood that a user will post an ad as she may consider it inappropriate to be charged a fee to respond to each of potentially many desirable suitors. Due to the direct relevance of this usability issue to this key business goal and the existence of clear potential solutions to address the problem, the issue was classified as high priority.

While frequency data may prove useful in prioritizing issues that are otherwise equivalent in terms of their nature and impact, such data is *not required*.

#### **QUALITATIVE RESEARCH AND THE EVALUATOR EFFECT**

The “evaluator effect” refers to the finding of large variance in both detection of usability issues and assignment of severity to those issues among usability professionals [8, 6, 16]. Given the compelling nature of this phenomenon and the existence of multiple corroborative studies, one may

conclude that it is inappropriate to base design decisions on qualitative data obtained from one participant that is reported by a single usability professional. Given the inherent lack of reliability in the reporting of qualitative usability findings, one might argue that quantitative data from numerous participants reported by multiple usability evaluators would be necessary to make meaningful (and valid) design recommendations. Despite the apparent cogency of this argument, we challenge it on several grounds.

While the above mentioned studies present percentages of reported usability issues, such issues are not presented in the framework of the Yahoo! Personals example above. For example, Jacobsen et al. [8] classified a behavior as a usability issue if it met one of various criteria including “the user expresses surprise” or “the user makes a design suggestion” (p. 1337). A more rigorous classification method (in which accounts of usability issues must include cause, impact, and potential solutions) may have reduced the number of usability issues reported in the above studies and increased consistency across evaluators [6].

Furthermore, when usability issues are presented as complete accounts of user behavior as part of a *formative* research project, the lack of overlap in reported usability findings becomes less important. However, as evidenced by the summary below provided by Hertzum and Jacobsen [6], the evaluator effect literature has not made clear the distinction between formative and summative usability research:

“Hence, it is highly questionable to use a thinking-aloud study with one evaluator as an authoritative statement about what problems an interface contains...The simplest way of coping with the evaluator effect, which cannot be completely eliminated, is to involve multiple evaluators in usability evaluations.” [6, p. 421]

We agree with the first sentence of the statement above which claims that one evaluator is likely insufficient to *assess* the quality of a system. However, we disagree with the associated recommendation that implies that multiple evaluators are required for a *formative* usability evaluation to be valid (or at least of any practical usefulness). As we have argued throughout, when the goal is one of *improvement* to products, this is not the case.

As long as an account of user behavior is reasonable and relevant to the business goals of product stakeholders, it is not necessary for that usability issue to overlap with those reported by other evaluators; each usability issue addressed will lead to an improved product. (While it is reasonable to assume that important issues may go undetected by a single evaluator, it does not invalidate the usability study methodology as per our detailed discussion earlier.)

Regarding assignment of severity in the above studies, the authors did not take into account the business goals of

product stakeholders. Asking a usability professional to assign severity to a usability issue without the benefit of business context is inappropriate, and it is not surprising that it would lead to erratic assignments of severity. In fact, lack of a proper understanding of the business context in which research is conducted may even naturally lead to variance in the *definition* of a “usability problem.”<sup>5</sup>

### WHY THE “HOW MANY USERS IS ENOUGH?” DEBATE HURTS THE USABILITY PROFESSION

In many organizations, the primary function of usability professionals is to improve the quality (and usability) of products and it is often challenging for such professionals to be heard among the many powerful organizations present in a company (such as product management, engineering, and business development). However, it has been argued that the problems faced by usability professionals are partly of their own making:

“Frankly, we think that UCD professionals too often shoot themselves in their own feet, *inadvertently behaving in ways that limit their influence* or keep them in the periphery.” [22, p. 19, emphasis added]

Usability professionals often rely on new research to propel their organizations forward and increase their influence. However, in our view, the “How many users is enough?” debate does little to achieve these aims. While valuable as an effort to improve the ability to assess the quality of products, it has drawn attention away from the key aims of usability professionals, namely to improve products and broaden their sphere of influence.

We agree with the common assumption that the need to improve products must be balanced with constraints such as cost and schedule [11, 26]. However, when the goal is to improve rather than assess products, concerns about cost and schedule should not dominate discussions of planned usability research. As argued above, we feel that each individual participant studied can potentially provide valuable insight that can lead to substantial improvement of products.

In this regard, other research advocating quantitative data to justify usability recommendations [9] is also misguided. The justification for usability recommendations should lie in the fidelity of the qualitative account provided for user difficulty and confusion and not in quantitative data (such as success rates). In fact, the widespread desire of usability professionals to find quantitative support to *validate* usability recommendations provides instant credibility to the misguided attacks on the validity of usability studies. While quantitative data can be helpful in prioritizing the importance of usability issues, it is completely unnecessary

to validate qualitative findings that provide a clear and cogent account of participant behavior.

Spool has argued that the usability profession is in a crisis as it cannot come to an agreement on the “basic elements of a quality testing protocol” [1, p. 699]. We agree that a crisis exists, but it is one of our own making in that we have inappropriately blurred the distinction between formative and summative research. Our proposed solution is to join the chorus who advocate focusing on improving products in practical settings as part of an iterative design process [2]. While debates regarding sample sizes may be productive, the usability profession would be better served by improving the ways in which it articulates the value and validity of formative research.

### ACKNOWLEDGMENTS

We thank Jeralyn Reese and Klaus Kaasgaard for their valuable input on earlier versions of this paper.

### REFERENCES

1. Barnum, C., Bevan, N., Cockton, G., Nielsen, J., Spool, J., Wixon, D. The “Magic number 5”: Is it enough for Web testing? In *Proc. CHI 2003*, ACM Press (2003), 698-699.
2. Becker, L. 90% of all usability testing is useless. (2004, June). <http://www.adaptivepath.com/publications/essays/archives/000328.php>.
3. Cockton, G., & Woolrych, A. Sale must end: Should discount methods be cleared off HCI’s shelves? *Interactions*, 9, 5 (2002), 13-18.
4. Dumas, J., & Redish, J. *A practical guide to usability testing*. Portland, OR: Intellect, 1993.
5. Gilmore, D. Understanding and overcoming resistance to ethnographic design research. *Interactions*, 9, 3 (2002), 29-35.
6. Hertzum, M. & Jacobsen, N. The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, 13 (2001), 421-443.
7. Hudson, W. HCI and the Web: How many users does it take to change a Web site? *SIGCHI Bulletin*, 33, 3 (2001), 6.
8. Jacobsen, N., Hertzum, M., & John, B. The evaluator effect in usability studies: Problem detection and severity judgments. In *Proc. Human Factors and Ergonomics Society 42<sup>nd</sup> Annual Meeting, 5-9 October 1998*, Santa Monica: Human Factors and Ergonomics Society (1998), 1336 – 1340.
9. James, J. How do we analyze and report valid usability results in a more timely and usable manner? (2002). [http://www.usabilityworks.net/idea\\_market\\_2002/timely\\_results\\_idea\\_mkt.htm](http://www.usabilityworks.net/idea_market_2002/timely_results_idea_mkt.htm).

---

<sup>5</sup> Thanks to Klaus Kaasgaard for pointing this out.

10. Kuniavsky, M. *Observing the user experience: A practitioner's guide to user research*. San Francisco, CA: Morgan Kaufmann, 2003.
11. Lewis, J. R. Problem discovery in usability studies: A model based on the binomial probability formula. In *Proc. Fifth International Conference on Human-Computer Interaction*. Orland, FL: Elsevier (1993), 666-671.
12. Lewis, J. R. Sample sizes for usability studies: Additional considerations. *Human Factors*, 36 (1994), 368-378.
13. Manning, H. Must the sale end? *Interactions*, 9, 6 (2002), 55-56.
14. Mayhew, D. *The usability engineering lifecycle: A practitioner's handbook for user interface design*. San Diego, CA: Academic Press, 1999.
15. Medlock, M. C., Wixon, D., Terrano, M., Romero, R. L., & Fulton, B. Using the RITE method to improve products: A definition and a case study. *Usability Professionals Association*, Orlando, FL, July 2002.
16. Molich, R., Ede, M., Kaasgaard, K., & Karyukin, B. Comparative usability evaluation. *Behaviour & Information Technology*, 23, 1 (2004), 65-74.
17. Nielsen, J. Estimating the number of subjects needed for a thinking aloud test. *International Journal of Human-Computer Studies*, 41, 3 (1994), 385-397.
18. Nielsen, J. Why you only need to test with 5 users. (March 19, 2000). <http://www.useit.com/alertbox>.
19. Nielsen, J. Success rate: The simplest usability metric. (2001, February). <http://www.useit.com/alertbox>.
20. Nielsen, J. Risks of quantitative studies. (2004, March). <http://www.useit.com/alertbox>.
21. Rubin, J. *Handbook of usability testing*. New York: Wiley & Sons, 1994.
22. Siegel, D., & Dray, S. Living on the edges: User-centered design and the dynamics of specialization in organizations. *Interactions*, 10, 5 (2003), 18-27.
23. Spool, J., & Schroeder, W. Testing Web sites: Five users is nowhere near enough. *Ext. Abstracts CHI 2001*, ACM Press (2001), 285-286.
24. Virzi, R. Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, 34, 4 (1992), 457-468.
25. Wixon, D. Evaluating usability methods: Why the current literature fails the practitioner. *Interactions*, 10, 4 (2003), 29-34.
26. Woolrych, A., & Cockton, G. Why and when five test users aren't enough. In *Proc. IHM-HCI, Vol. 2*, Cepadeus Editions (2001), 105-108.