MEASURING THE USER EXPERIENCE

Collecting, Analyzing, and Presenting Usability Metrics

TOM TULLIS
BILL ALBERT

MK
MORGAN KAUFMANN

CHAPTER

Issues-Based Metrics

5

Most usability professionals probably consider identifying usability issues and providing design recommendations the most important part of their job. A usability issue might involve confusion around a particular term or piece of content, method of navigation, or just not noticing something that should be noticed. These types of issues, and many others, are typically identified as part of an iterative process in which designs are being evaluated and improved. This process provides tremendous value to product design and is the cornerstone of the usability profession.

Usability issues are generally thought of as purely qualitative. They typically include a description of a problem one or more participants experienced and, in many cases, an assessment of the underlying cause of the problem. Most usability professionals also include specific recommendations for remedying the problem, and some usability professionals report positive findings (i.e., something that worked particularly well).

Most usability professionals don't strongly associate metrics with usability issues. This may be because of the gray areas in identifying issues or because identifying issues is part of an iterative design process, and metrics are perceived as adding no value. However, not only is it possible to measure usability issues, but doing so also adds value in product design while not slowing down the iterative process.

In this chapter we review some simple metrics around usability issues. We will also discuss different ways of identifying usability issues, prioritizing the importance of different types of issues, and factors you need to think about when measuring usability issues.

5.1 IDENTIFYING USABILITY ISSUES

Identifying usability issues can be easy or difficult and everything in between. Sometimes the problems that participants encounter are so obvious that the issues practically punch you in the face. Other times usability issues are much more subtle and require careful observation. At the highest level, three things help to identify usability issues:

Understanding what is (and isn't) an issue. The more issues you've observed over the course of your professional career, the easier it is to know which issues are real and which are not. This is the value of being a usability expert.

Knowing both the product and the usability questions that arise. The better you know the product, the easier it is to know when someone has trouble with it. When you know what to look for, spotting an issue is much easier.

Being very observant of participant behavior. Listen carefully with an open mind and observe different types of behavior, such as facial expressions and body language. The more you pay attention, the easier it is to spot a usability issue.

5.2 WHAT IS A USABILITY ISSUE?

What do we mean by usability issues? There's no simple definition, so perhaps the best way to characterize them is to give some examples:

- Anything that prevents task completion
- Anything that takes someone ''off-course''
- Anything that creates some level of confusion
- Anything that produces an error
- Not seeing something that should be noticed
- Assuming something is correct when it is not
- Assuming a task is complete when it is not
- Performing the wrong action
- Misinterpreting some piece of content
- Not understanding the navigation

A key point to consider in defining usability issues is how they will be used. The most common use is in an iterative design process focused on improving the product. In that context, the most useful issues are those that point to possible improvements in the product. In other words, it helps if issues are reasonably actionable. If they don't directly point to a part of the interface that was causing a problem, they should at least give you some hint of where to begin looking. For example, we once saw an issue in a usability test report that said, ''The mental model of the application does not match the user's mental model.'' And that was it. Although this may be an interesting observation in some theoretical sense, it does very little to guide designers and developers in addressing the issue.

On the other hand, consider an issue like this: ''Many participants were confused by the top-level navigation menu, often jumping around from one section to another, trying to find what they were looking for.'' Particularly if this issue is followed by a variety of detailed examples describing what happened, it could be very helpful. It tells you where to start looking (the top-level navigation), and the more detailed examples may help focus on some possible solutions. Molich, Jeffries, and Dumas (2007) conducted an interesting study of usability recommendations and ways to make them more useful and usable.

Of course, not all usability issues should be avoided. Some usability issues are positive, such as aspects of a product that exceed a user's expectation for ease of use, efficiency, or satisfaction. These are sometimes called usability "findings," since the term *issues* often has negative connotations. Here are some examples of positive findings:

- Supporting the user in completing a complex transaction without any confusion and in the most efficient way possible
- Anticipating a user's needs at every step of a process
- Educating a user without any effort involved
- Displaying complex information in a clear, simple format that users can easily understand

The main reason for reporting positive findings, in addition to providing some positive reinforcement for the project team, is to make sure that these aspects of the interface don't get "broken" in future design iterations.

5.2.1 Real Issues versus False Issues

One of the most difficult parts of any usability professional's job is determining which usability issues are real and which are merely an aberration. Obvious issues are those that most, if not at all, participants encounter. For example, it may be obvious when participants select the wrong option from a poorly worded menu, get taken down the wrong path, and then spend a significant amount of time looking for their target in the wrong part of the application. These are the "no-brainer" issues that are easy for almost anyone to identify.

Some usability issues are much less obvious, or it's not completely clear whether something is a real issue. For example, what if only 1 out of 10 participants expresses some confusion around a specific piece of content or terminology on a website? Or if only 1 out of 12 participants doesn't notice something she should have? At some point the usability specialist must decide whether what he observed is likely to be repeatable with a larger population. In these situations, ask whether the participant's behavior, thought process, perception, or decisions during the task were *logical*. In other words, is there a consistent story or reasoning behind her actions or thoughts? If so, then it may be an issue even if only one participant encountered it. On the other hand, no apparent rhyme or reason behind the behavior may be evident. If the participant can't explain why he did what he did, and it only happened once, then it's likely to be idiosyncratic and should probably be ignored.

For example, assume that you observed one participant click on a link on a web page that started him down the wrong path for accomplishing the task. At the end of the task, you might ask him why he clicked on that link. If he says that he clicked on it simply because it was there in front of him, you might discount this as a false issue. On the other hand, if he says that the wording of the link made it seem like a reasonable place to begin the task, it may be a genuine issue.

5.3 HOW TO IDENTIFY AN ISSUE

The most common way to identify usability issues is during a study in which you are directly interacting with a participant. This might be in person or over the phone using remote testing technology. A less common way to identify usability issues is through some automated techniques such as an online study. This is where you don't have an opportunity to directly observe participants but only have access to their behavioral and self-reported data. Identifying issues through this type of data is more challenging but still quite possible.

Possible usability issues might be predicted beforehand and tracked during test sessions. But be careful that you're really *seeing* the issues and not just finding them because you expected to. Your job is certainly easier when you know what to look for, but you might also miss other issues that you never considered. In our testing, we typically have an idea of what to look for, but we also try to keep an open mind to spot the surprise issues. There's no ''right'' approach; it all depends on the goals of the evaluation. When evaluating products that are in an early conceptual stage, it's more likely that you don't have preset ideas about what the usability issues are. As the product is further refined, you may have a clearer idea of what specific issues you're looking for.

THE ISSUES YOU EXPECT MAY NOT BE THE ONES YOU FIND

One of the earliest sets of guidelines for designing software interfaces was published by Apple (1982). It was called the *Apple IIe Design Guidelines,* and it contained a fascinating story of an early series of usability tests Apple conducted. They were working on the design of a program called *Apple Presents Apple*, which was a demonstration program for customers to use in computer stores. One part of the interface to which the designers paid little attention was asking users whether their monitor was monochrome or color. The initial design of the question was "Are you using a black-and-white monitor?" (They had predicted that users might have trouble with the word *monochrome*.) In the first usability test, they found that a majority of the participants who used a monochrome monitor answered this question incorrectly because their monitor actually displayed text in green, not white!

What followed was a series of hilarious iterations involving questions such as "Does your monitor display multiple colors?" or "Do you see more than one color on the screen?"—all of which kept failing for some participants. In desperation, they were considering including a developer with every computer just to answer this question, but then they finally hit on a question that worked: "Do the words above appear in several different colors?" In short, the issues you expect may not be the issues you find.

5.3.1 In-Person Studies

The best way to facilitate identifying usability issues during an in-person study is using a think-aloud protocol. This involves having participants verbalize their thoughts as they are working through the tasks. Typically, the participants are reporting what they are doing, what they are trying to accomplish, how confident they are about their decisions, their expectations, and why they performed certain actions. Essentially, it's a stream of consciousness focusing on their interaction with the product. During a think-aloud protocol, you might observe the following:

- Verbal expressions of confusion, frustration, dissatisfaction, pleasure, or surprise
- Verbal expressions of confidence or indecision about a particular action that may be right or wrong
- Participants *not* saying or doing something that they should have done or said
- Nonverbal behaviors such as facial expressions and/or eye movements

Any of these might point to usability issues.

5.3.2 Automated Studies

Identifying usability issues through automated studies requires careful data collection. The key is to allow participants to enter verbatim comments at a page or task level. In most automated studies, several data points are collected for each task: success, time, ease-of-use rating, and verbatim comments. The verbatim comments are the best way to understand any possible issues.

One way to collect verbatim comments is to require the participant to provide a comment at the conclusion of each task. This might yield some interesting results, but it doesn't always yield the best results. An alternative that seems to work better is to make the verbatim comment conditional. If the participant provides a low ease-of-use score (e.g., not one of the two highest ratings), then she is asked to provide feedback about why she rated the task that way. Having a more pointed question usually yields more specific, actionable comments. For example, participants might say that they were confused about a particular term or that they couldn't find the link they wanted on a certain page. This type of task-level feedback is usually more valuable than one question after they complete all the tasks (post-study).

5.3.3 When Issues Begin and End

When does a usability issue begin? It might start with some initial confusion or deviation away from ideal behavior. For example, a participant may express some question about a link on a website, click on it, and then get taken down a black hole. In this case, the issue really started when the question or doubt first arose within the participant's mind. In another situation, a participant may voice some confusion about what button to press on a remote control. In many cases the participant is not even aware that he is experiencing an issue until the very end, when he discovers that he didn't get the answer or result he was looking for. For example, a user might be going along a series of steps to install some software, only to realize at the end of the process that she made a mistake. In this case, the issue began at the point of departure from the ideal process.

The next obvious question is, when does an issue end? There are a few situations that typically indicate the end of an issue. In one common situation, an issue ends with task failure. The participant either realizes that she got the wrong answer or gives up. This is when the participant throws up her hands and the facilitator says, "Okay, let's move on to the next task." Other usability issues end when another one begins—in other words, a different problem arises and takes the place of the original issue. The new problem is sufficiently different from the original problem, and it has become the driving force behind her current behavior. Finally, in some situations, the participant recovers from the issue. This is when a participant says, "Oh, *now* I get it!" They identified the source of the confusion and might even go back several steps and show you where they were confused and why. This is always enlightening, since it gives you direct access to their thoughts about what originally precipitated the issue. An issue might also end when it's no longer relevant. The participant either recovered in some way or it just became insignificant.

5.3.4 Granularity

One of the most important decisions when identifying usability issues is how detailed or granular they should be. High granularity is useful for identifying very specific problems but makes it easy to miss the big picture. This level of detail usually focuses on one particular behavior at a specific point in time. For example, a user might fail to notice a particular link on a web page or might misinterpret a specific field label on a form. In either case, the issue is described in a fairly detailed way.

When you're looking for usability issues with less granularity, it's easier to see the big picture. For example, you might state that a user often jumped around from one section of the site to another in search of information and was obviously confused about the overall organization of the site. Or maybe the information density was so high on many of the pages that the user missed many of the links.

Identifying issues at a granular level or at a high level can both be appropriate and useful. Just keep in mind the audience and the goals of the study. If the product is at an early stage of design, capturing the big issues is probably more important. If the design is fairly well developed, then a more detailed examination of the issues may be more helpful.

5.3.5 Multiple Observers

There is power in numbers. Whenever possible, try to have multiple observers on hand to identify usability issues—ideally at least two or three observers. This is particularly true when you are the one moderating the sessions. Things happen quickly in a usability session, and the moderator must be focused on the participant rather than on taking notes. Some studies have shown that more observers identify more usability issues than fewer observers (Hertzum, Jacobsen, & Molich, 2002; Molich et al., 1998). Of course, there are some situations where it's not practical to have more than one observer. In this case, it's helpful to review videos of the sessions to capture any issues that you might have originally missed. This is particularly important when you're a one-person team. If you're evaluating a product alone and you don't have access to audio or video recordings, make sure you take good notes!

5.4 SEVERITY RATINGS

Not all usability issues are the same: Some are more serious than others. Some usability issues annoy or frustrate users, whereas others cause them to make the wrong decisions or lose data. Obviously, these two different types of usability issues have a very different impact on the user experience, and severity ratings are a useful way to deal with them.

Severity ratings help focus attention on the issues that really matter. There's nothing more frustrating for a developer or business analyst than being handed a list of 82 usability issues that all need to be fixed immediately. By prioritizing usability issues, you're much more likely to have a positive impact on the design, not to mention lessening the likelihood of making enemies with the rest of the design and development team.

The severity of usability issues can be classified in many ways, but most severity rating systems can be boiled down to two different types. In one type of rating system, severity is based purely on the impact on the user experience: The worse the user experience, the higher the severity rating. A second type of severity rating system tries to bring in multiple dimensions or factors. These dimensions usually include impact on the user experience, predicted frequency of use, and impact on the business goals. In this case, the highest severity rating would mean that the participant failed a task that is frequently done and is critical to the business. Think of someone who wants to buy something on a website, finds the product, and then fails at the actual purchase!

5.4.1 Severity Ratings Based on the User Experience

Many severity ratings are based solely on the impact on the user experience. These rating systems are easy to implement and provide very useful information. They usually have between three and five levels—often something like low, medium, and high severity. In some rating systems there is a "catastrophe" level, which is essentially a showstopper (delaying product launch or release—Nielsen, 1993). This, of course, is above and beyond a "high," which is normally reserved for the biggest usability issues but might not necessarily delay the launch of the product. Other severity rating systems have a ‘‘cosmetic’’ rating (Nielsen, 1993; Rubin, 1994). These are superficial usability issues that can be addressed if time and budget allow. For example, some participants might comment on the colors or other aspects of the visual design.

Wilson (1999) proposes a five-level rating system: level 5 (minimal error), level 4 (minor but irritating), level 3 (moderate: waste of time but no loss of data), level 2 (severe problem causing loss of data), and level 1 (catastrophic error, causing loss of data or damage to hardware/software). He also suggests that your severity rating system should be consistent with the bug-tracking system used in your organization. This will help with adoption of the system, as well as with tracking and fixing each of the issues.

When choosing a severity rating system, it’s important to look at your organization and the product you are evaluating. Often, a three-level system works well in many situations:

Low: Any issue that annoys or frustrates participants but does not play a role in task failure. These are the types of issues that may lead someone off course, but he still recovers and completes the task. This issue may only reduce efficiency and/or satisfaction a small amount, if any.

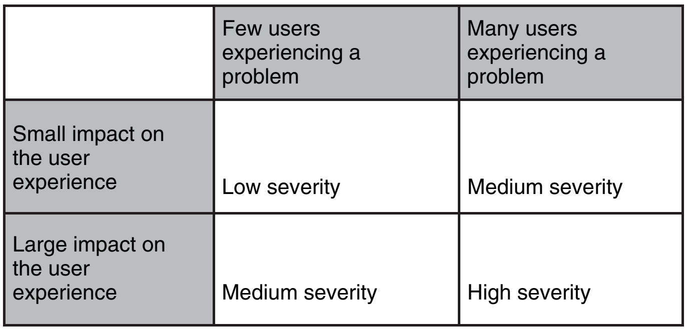
Medium: Any issue that contributes to but does not directly prevent task failure. Participants often develop workarounds to get to what they need. These issues have an impact on effectiveness and most likely efficiency and satisfaction.

High: Any issue that directly leads to task failure. Basically, there is no way to encounter this issue and still complete the task. This type of issue has a significant impact on effectiveness, efficiency, and satisfaction.

5.4.2 Severity Ratings Based on a Combination of Factors

Severity rating systems that use a combination of factors usually are based on the impact on the user experience coupled with frequency of use and/or impact on the business goals. Nielsen (1993) provides an easy way to combine the impact on the user experience and frequency of use on severity ratings (Figure 5.1). This severity rating system is intuitive and easy to explain.

Rubin (1994) offers a different way of looking at the combination of severity and the frequency of occurrence of issues. First, he assigns a *severity rating* on a 4-point scale (1 = irritant, 2 = moderate, 3 = severe, 4 = unusable). Next, he assigns a *frequency of occurrence,* also on a 4-point scale (1 = occurs < 10 percent of the time; 2 = occurs 11 to 50 percent of the time; 3 = occurs 51 to 89 percent of the time; 4 = occurs more than 90 percent of the time). He then simply adds the two scores to arrive at a criticality score between 2 and 8. This approach gives a numeric severity score that may be helpful when combined with other types of data.



	Few users experiencing a problem	Many users experiencing a problem
Small impact on the user experience	Low severity	Medium severity
Large impact on the user experience	Medium severity	High severity

FIGURE 5.1

Severity rating scale taking into account problem frequency and impact on the user experience. *Source*: Adapted from Nielsen (1993).

Building on Rubin's method for combining different types of scores, it's possible to add a third dimension based on *importance to the business goals*. For example, you might combine three different 3-point scales:

- Impact on the user experience (1 = low, 2 = medium, 3 = high)
- Predicted frequency of occurrence (1 = low, 2 = medium, 3 = high)
- Impact on the business goals (1 = low, 2 = medium, 3 = high)

By adding up the three scores, you now have an overall severity rating ranging from 3 to 9. Of course, a certain amount of guesswork is involved in coming up with the levels, but at least all three factors are being taken into consideration.

5.4.3 Using a Severity Rating System

Once you have settled upon a severity rating system, you still need to consider a few more things. First, be consistent: Decide on one severity rating system, and use it for all your studies. By using the same severity rating system, you will be able to make meaningful comparisons across studies, as well as help train your audience on the differences between the severity levels. The more your audience internalizes the system, the more persuasive you will be in promoting design solutions.

Second, clearly communicate what each level means. Provide examples of each level as much as possible. This is particularly important for other usability specialists on your team who might also be assigning ratings. It's important that developers, designers, and business analysts understand each severity level. The more the "nonusability" audience understands each level, the easier it will be to influence design solutions for the highest-priority issues.

Third, try to have more than one usability specialist assign severity ratings to each issue. One approach that works well is to have the usability specialists independently assign severity ratings to each of the issues, then discuss any of the issues where they gave different ratings and try to agree on the appropriate level.

Finally, there's some debate about whether usability issues should be tracked as part of a larger bug-tracking system (Wilson & Coyne, 2001). Wilson argues that it is essential to track usability issues as part of a bug-tracking system because it makes the usability issues more visible, lends more credibility to the usability team, and makes it more likely that the issues will be remedied. Coyne suggests that usability issues, and the methods to fix them, are much more complex than typical bugs. Therefore, it makes more sense to track usability issues in a separate database. Either way, it's important to track the usability issues and make sure they are addressed.

5.4.4 Some Caveats about Severity Ratings

Not everyone believes in severity ratings. Kuniavsky (2003) suggests letting your audience provide their own severity ratings. He argues that only those who are deeply familiar with the business model will be able to determine the relative priority of each usability issue.

Bailey (2005) strongly argues against severity rating systems altogether. He cites several studies that show there is very little agreement between usability specialists on the severity rating for any given usability issue (Catani & Biers, 1998; Cockton & Woolrych, 2001; Jacobsen, Hertzum, & John, 1998). All of these studies generally show that there is very little overlap in what different usability specialists identify as a high-severity issue. Obviously, this is troubling given that many important decisions may be based on severity ratings.

Hertzum et al. (2002) highlight a potentially different problem in assigning severity ratings. In their research they found that when multiple usability specialists are working as part of the same team, each usability specialist rates the issues she personally identifies as more severe than issues identified by the other usability specialists on their own team. This is known as an evaluator effect, and it poses a significant problem in relying on severity ratings by a single usability professional.

So where does this leave us? We believe that severity ratings are not perfect, but they still serve a useful purpose. They help direct attention to at least some of the most pressing needs. Without severity ratings, the designers or developers will simply make their own priority list, perhaps based on what's easiest to implement. Even though there is some subjectivity involved in assigning severity ratings, they're better than nothing. Most key stakeholders understand that there is more art than science involved, and they interpret the severity ratings within this broader context.

5.5 ANALYZING AND REPORTING METRICS FOR USABILITY ISSUES

Once you've identified and prioritized the usability issues, it's helpful to do some analyses of the issues themselves. This lets you derive some metrics related to the usability issues. Exactly how you do this will largely depend on the type of usability questions you have in mind. Three general questions can be answered by looking at metrics related to usability issues:

- How is the overall usability of the product? This is helpful if you simply want to get an overall sense of how the product did.
- Is the usability improving with each design iteration? Focus on this question when you need to know how the usability is changing with each new design iteration.
- Where should I focus my efforts to improve the design? The answer to this question is useful when you need to decide where to focus your resources.

All of the analyses we will examine can be done with or without severity ratings. Severity ratings simply add a way to filter the issues. Sometimes it's helpful to focus on the high-severity issues. Other times it might make more sense to treat all the usability issues equally.

5.5.1 Frequency of Unique Issues

The most common way to measure usability issues is to simply count the unique issues. Analyzing the frequency of unique issues is most useful in an iterative design process when you want some basic data about how the usability is changing with each iteration. For example, you might observe that the number of issues decreased from 24 to 12 to 4 through the first three design iterations. These data are obviously trending in the right direction, but they're not necessarily iron-clad evidence that the design is significantly better. Perhaps the four remaining issues are so much bigger than all the rest that without addressing them, everything else is unimportant. Therefore, we suggest a thorough analysis and explanation of the issues when presenting this type of data.

Keep in mind that this frequency represents the number of *unique issues*, not the *total number of issues* encountered by all participants. For example, assume Participant A encountered ten issues, whereas Participant B encountered 14 issues, but 6 of those issues were the same as those from Participant A. If A and B were the only participants, the total number of unique issues would be 18 (10 + 14 − 6). Figure 5.2 shows an example of how to present the frequency of usability issues when comparing more than one design.

The same type of analysis can be performed using usability issues that have been assigned a severity rating. For example, if you have classified your usability issues into three levels (low, medium, and high severity), you can easily look at the number of issues by each type of severity rating. Certainly the most telling data item would be the change in the number of high-priority issues with each design iteration. Looking at the frequency of usability issues by severity rating, as illustrated in Figure 5.3, can be very informative since it is an indicator of whether the design effort between each iteration is addressing the most important usability issues.

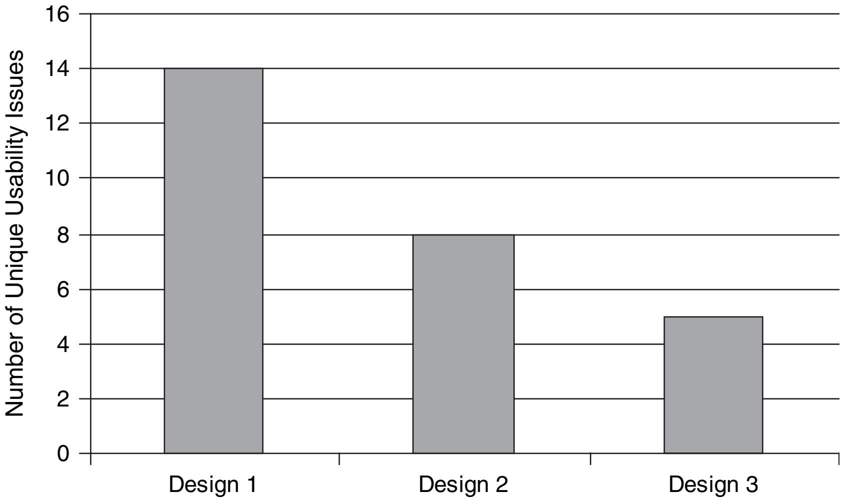


FIGURE 5.2

Example data showing the number of unique usability issues by design iteration. Ideally, the number of issues decreases with each new design iteration.

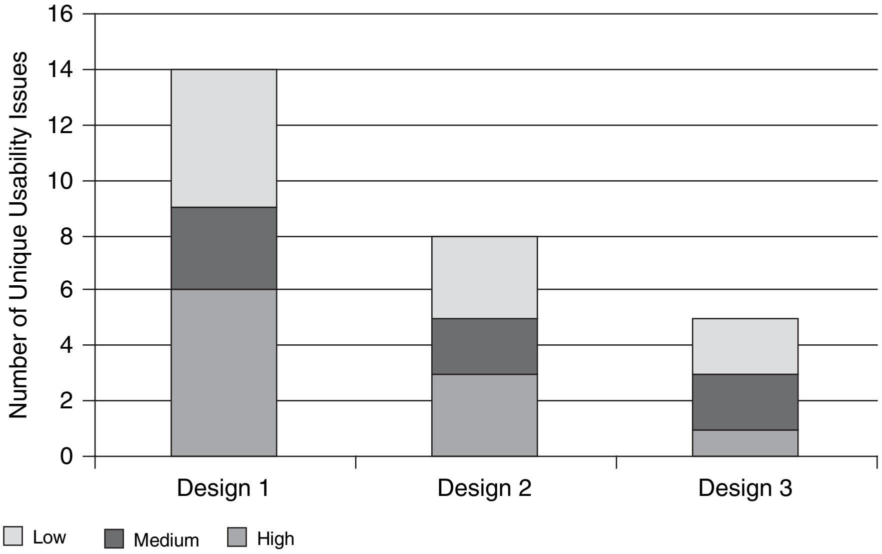


FIGURE 5.3

Example data showing the number of unique usability issues by design iteration, categorized by severity rating. The change in the number of high-severity issues is probably of key interest.

5.5.2 Frequency of Issues per Participant

It can also be informative to look at the number of issues each participant encountered. Over a series of design iterations, you would expect to see this number decreasing along with the total number of unique issues. For example, Figure 5.4 shows the average number of issues encountered by each participant for three design iterations.

Of course, this analysis could also include the average number of issues per participant broken down by severity level. If the average number of issues per participant is not declining over a series of iterations, but the total number of unique issues is declining, then you know there is more consistency in the issues that the participants are encountering. This would indicate that the issues encountered by fewer participants are being fixed whereas those encountered by more participants are not.

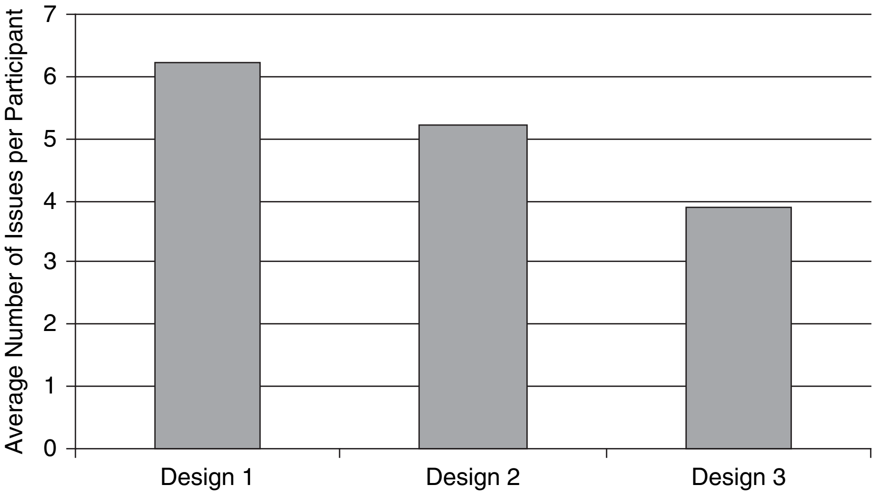


FIGURE 5.4

Example data showing the average number of usability issues encountered by participants in each of three usability tests.

5.5.3 Frequency of Participants

Another useful way to analyze usability issues is to observe the frequency or percentage of participants who encountered a specific issue. For example, you might be interested in whether participants correctly used some new type of navigation element on your website. You report that half of the participants encountered a specific issue in the first design iteration, and only one out of ten encountered the same issue in the second design iteration. This is a useful metric

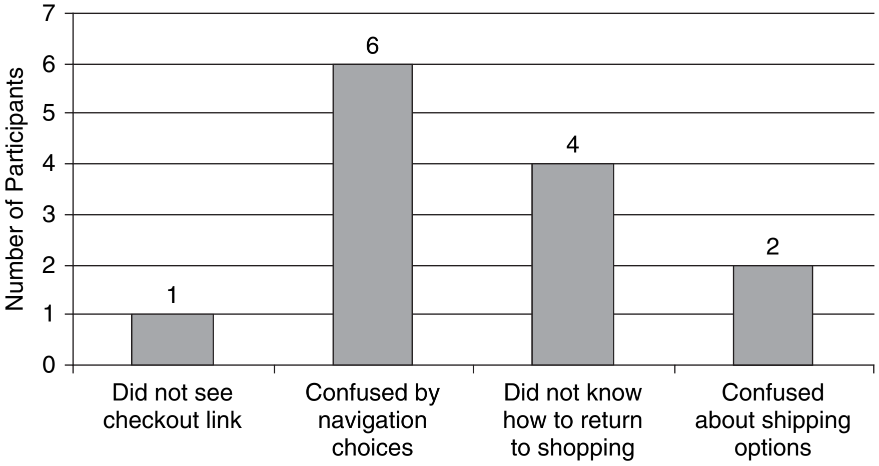


FIGURE 5.5

Example data showing the frequency of participants who experienced specific usability issues.

when you need to focus on whether you are improving the usability of specific design elements as opposed to making overall usability improvements.

With this type of analysis, it's important that your criteria for identifying specific issues are consistent between participants and designs. If a description of a specific issue is a bit fuzzy, your data won't mean very much. It's a good idea to explicitly document the issue's exact nature, thereby reducing any interpretation errors across participants or designs. Figure 5.5 shows an example of this type of analysis.

The use of severity ratings with this type of analysis is useful in a couple of ways. First, you could use the severity ratings to focus your analysis only on the high-priority issues. For example, you could report that there are five outstanding high-priority usability issues. Furthermore, the percentage of participants who are experiencing these issues is decreasing with each design iteration. Another form of analysis is to aggregate all the high-priority issues to report the percentage of participants who experienced any high-priority issue. This helps you to see how overall usability is changing with each design iteration, but it is less helpful in determining whether to address a specific usability problem.

5.5.4 Issues by Category

Sometimes it's helpful to know where to focus design improvements from a tactical perspective. Perhaps you feel that only certain areas of the product are causing the most usability issues, such as navigation, content, terminology, and so forth. In this situation, it can be useful to aggregate usability issues into broad categories. Simply examine each issue and then categorize it into a type of issue. Then look at the frequencies of issues that fall into each category. Issues can be

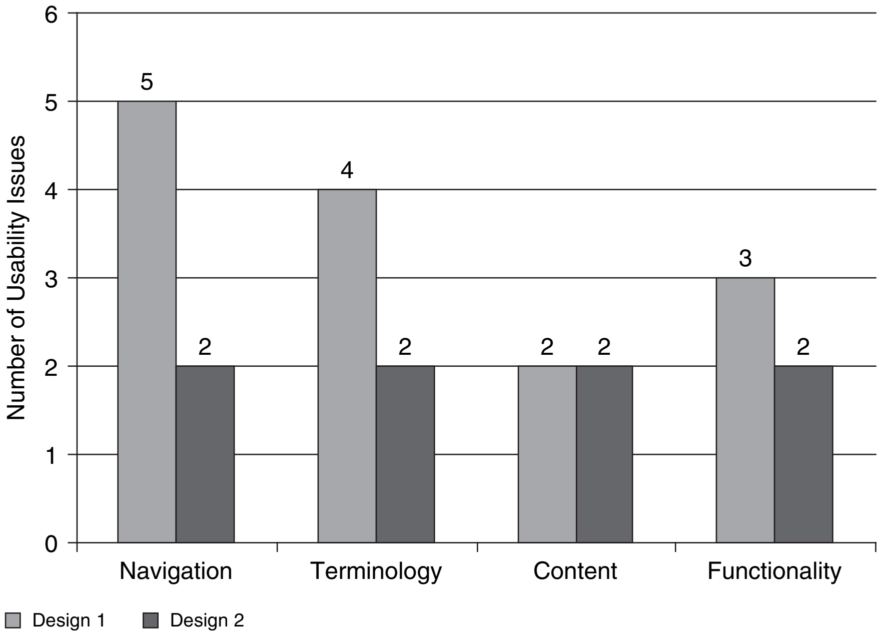


FIGURE 5.6

Example data showing the frequency of usability issues categorized by type. Notice that both navigation and terminology issues were improved from the first to the second design iteration.

categorized in many different ways. Just make sure the categorization makes sense to you and your audience, and use a limited number of categories, typically three to eight. If there are too many categories, it won't provide much direction. Figure 5.6 provides an example of usability issues analyzed by category.

5.5.5 Issues by Task

Issues can also be analyzed at a task level. You might be interested in which tasks lead to the most issues, and you can report the number of unique issues that occur for each. This will identify the tasks you should focus on for the next design iteration. Alternatively, you could report the frequency of participants who encounter any issue for each task. This will tell you the pervasiveness of a particular issue. The greater the number of issues for each task, the greater the concern should be.

If you have assigned a severity rating to each issue, it might be useful to analyze the frequency of high-priority issues by task. This is particularly effective if you want to focus on a few of the biggest problems and your design efforts are oriented toward specific tasks. This is also helpful if you are comparing different design iterations using the same tasks.

5.5.6 Reporting Positive Issues

It's important to report positive issues (e.g., Dumas et al., 2004) for the following reasons:

- Designers and developers should be able to consider you an ally. If you go to them only with problems, it's hard to build any goodwill. In fact, it's better to begin every usability report with at least one positive issue. Starting with a negative issue often sets a somber tone for the rest of the report or presentation.
- You should document positive issues so they can be propagated to other areas of the product as appropriate. It's helpful to know what's working well in the design and to identify other areas that could benefit from a similar solution. It's also important not to mess up the positive aspects across design iterations!
- Reporting positive issues enhances your credibility. It's critical that you are, and appear to be, neutral. If you only report problems, some may question your neutrality.

The analysis and presentation of positive issues are essentially the same as for any other issue. The only unique form of analysis involving positive issues is to calculate a ratio of positive to negative issues. It's questionable whether this is really useful, but some audiences may be interested in it.

5.6 CONSISTENCY IN IDENTIFYING USABILITY ISSUES

Much has been written about consistency and bias in identifying and prioritizing usability issues. Unfortunately, the news is not so good. Much of the research shows that there is very little agreement on what a usability issue is or how important it is.

Perhaps the most exhaustive set of studies, called CUE (Comparative Usability Evaluation) was coordinated by Rolf Molich. To date, six separate CUE studies have been conducted, dating back to 1998. Each study was set up in a similar manner. Different teams of usability experts all evaluated the same design. Each team reported their findings, including the identification of the usability issues along with their design recommendations. The first study, CUE-1 (Molich et al., 1998), showed very little overlap in the issues identified. In fact, only 1 out of the 141 issues was identified by all four teams participating in the study, and 128 out of the 141 issues were identified by single teams. Several years later, in CUE-2, the results were no more encouraging: 75 percent of all the issues were reported by only 1 of 9 usability teams (Molich & Dumas, 2006). CUE-4 (Molich & Dumas, 2006) showed similar results: 60 percent of all the issues were identified by only 1 of the 17 different teams participating in the study.

These data certainly do not bode well for the usability profession, but could there be alternative explanations? Some usability professionals have criticized the format of the CUE studies. For example, some have suggested that the teams differed dramatically in terms of professional experience. Others have claimed that the analysis of whether two issues from separate reports were really the same basic issue was too strict: that the criterion for calling two issues the same was set too high.

Tullis (2005) formed two different usability teams from a group of usability professionals who all work together and presumably have relatively consistent usability testing methods. Both teams independently evaluated the same website using a traditional lab test with real end-users (as opposed to an expert evaluation). He found that 38 percent of the usability issues were reported by both teams, 32 percent by one team, and 30 percent by the other team (Figure 5.7). Each team also prioritized the issues (low, medium, and high).

When looking only at the high-severity issues, 88 percent were shared among both teams (Figure 5.8). This is much more encouraging, since it suggests that perhaps many of the issues uniquely identified by each of the teams were the less important ones. Perhaps the bigger the issue, the more likely that it will be observed by different teams.

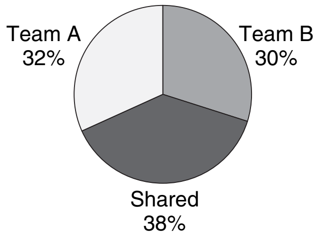


FIGURE 5.7

Data showing that 38 percent of the unique usability issues were identified by both teams.

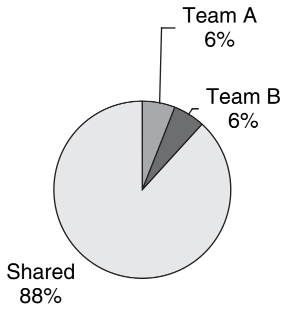


FIGURE 5.8

Data showing that 88 percent of the high-priority usability issues were identified by both teams.

5.7 BIAS IN IDENTIFYING USABILITY ISSUES

Many different factors can influence how usability issues are identified. Carolyn Snyder (2006) provides a review of many of the ways usability findings might be biased. She concludes that bias cannot be eliminated, but it must be understood. In other words, even though our methods have flaws, they are still useful.

We've distilled the different sources of bias in a usability study into six general categories:

Participants: Your participants are critical. Every participant brings a certain level of technical expertise, domain knowledge, and motivation. Some participants may be well targeted, and others may not. Some participants are comfortable in a lab setting, whereas others are not. All of these factors make a big difference in what usability issues you end up discovering.

Tasks: The tasks you choose have a tremendous impact on what issues are identified. Some tasks might be well defined with a clear end-state, whereas others might be open-ended, and yet others might be self-generated by each participant. The tasks basically determine what areas of the product are exercised and the ways in which they are exercised. Particularly with a complex product, this can have a major impact on what issues are uncovered.

Method: The method of evaluation is critical. Methods might include traditional lab testing or some type of expert review. Other decisions you make are also important, such as how long each session lasts, whether the participant thinks aloud, or how and when you probe.

Artifact: The nature of the prototype or product you are evaluating has a huge impact on your findings. The type of interaction will vary tremendously whether it is a paper prototype, functional or semifunctional prototype, or production system.

Environment: The physical environment also plays a role. The environment might involve direct interaction with the participant, indirect interaction via a conference call or behind a one-way mirror, or even at someone's home. Other characteristics of the physical environment, such as lighting, seating, observers behind a one-way mirror, and videotaping, can all have an impact on the findings.

Moderators: Different moderators will also influence the issues that are observed. A usability professional's experience, domain knowledge, and motivation all play a key role.

An interesting study that sheds some light on these sources of bias was conducted by Lindgaard and Chattratichart (2007). They analyzed the reports from the nine teams in CUE-4 who conducted actual usability tests with real users. They looked at the number of participants in each test, the number of tasks used, and the number of usability issues reported. The number of participants ranged from 5 to 15, the

AN UNEXPECTED WAY OF BIASING PARTICIPANTS' BEHAVIOR

One of our funniest experiences in learning how subtle aspects of the environment can impact the identification of usability issues involved a test of a prototype web application. Due to the availability of representative users, we had to set up a makeshift usability lab in a conference room at the participant's location. The participant sat at the conference table along with the moderator. Several of the project team members wanted to observe the sessions, so we agreed to let them sit along the wall as long as they promised not to say anything.

Everything was going fine, but after a while we noticed that some of the participants were showing a similar pattern of behavior: They would reach a page of the prototype that they somehow decided was the right page for their task, but then they couldn't figure out how to complete the task. In fact, it was the right page. Then we figured out what was happening: As the participant was navigating around the prototype trying to find the right page for a task, some of the observers would lean forward, very intently observing. When the participant reached the right page, the observers would lean back in relief, assuming he was almost done. Some of the participants picked up on this, using it as a cue that they had reached the right page! Once we glued the observers to their chairs (not really), this stopped, and we learned that the overriding issue was finding the right page!

number of tasks from 5 to 14, and the number of issues from 20 to 50. Looking across all the reports, a total of 106 unique usability problems categorized as serious or critical were identified. They found no significant correlation between the number of *participants* in the test and the percentage of usability problems found.

On the other hand, they did find a significant correlation between the number of *tasks* used and the percentage of usability problems found ($r = 0.82, p < 0.01$). When looking at the percentage of *new* problems uncovered, the correlation with the number of tasks was even higher ($r = 0.89$, $p < 0.005$). As Lindgaard and Chattratichart concluded, these results suggest "that with careful participant recruitment, investing in wide task coverage is more fruitful than increasing the number of users."

One technique that works well to increase the task coverage in a usability test is to define a core set of tasks that all participants must complete and another set that is derived for each participant. These additional tasks might be selected based on characteristics of the participant (e.g., an existing customer or a prospect), or they might be selected at random. Care must be exercised when making comparisons across participants, since not all participants had the same tasks. In this situation, you may want to limit certain analyses to the core tasks.

5.8 NUMBER OF PARTICIPANTS

There has been much debate about how many participants are needed in a usability test to reliably identify usability issues. (See Barnum et al., 2003, for a summary of the debate.) Nearly every usability professional seems to have an opinion. Not only are many different opinions floating around out there, but quite a few compelling studies have been conducted on this very topic. From this research, two different camps have emerged: those who believe that five participants is enough to identify most of the usability issues and those who believe that five is nowhere near enough.

5.8.1 Five Participants Is Enough

One camp believes that a majority, or about 80 percent, of usability issues will be observed with the first five participants (Lewis, 1994; Nielsen & Landauer, 1993; Virzi, 1992). This is known as the "magic number 5." One of the most important ways to figure out how many participants are needed in a usability test is to measure *p,* or the probability of a usability issue being detected by a single test participant. It's important to note that this *p* is different from the *p*-value that is used in tests of significance. The probabilities vary from study to study, but they tend to average around 0.3, or 30 percent. (See Turner, Nielsen, and Lewis, 2002, for a review of different studies.) In the seminal paper, Nielsen and Landauer (1993) found an average probability of 31 percent based on 11 different studies. This basically means that with each participant, about 31 percent of the usability problems are being observed.

Figure 5.9 shows how many issues are observed as a function of the number of participants when the probability of detection is 30 percent. (Notice that this assumes all issues have an equal probability of detection, which may be a big

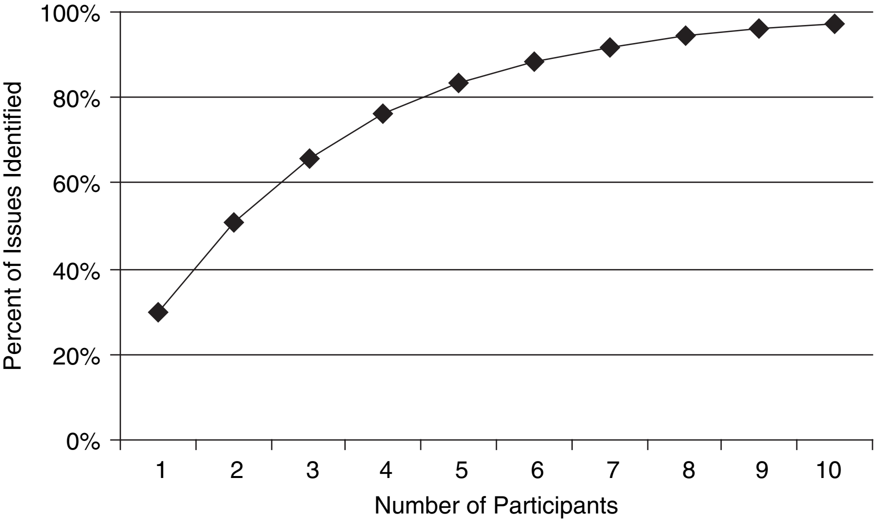


FIGURE 5.9

Example showing how many users are required to observe the total number of issues in a usability study, given a probability of detection equal to 30 percent with each participant.

assumption.) As you can see, after the first participant, 30 percent of the problems are detected; after the third participant, about 66 percent of the problems are observed; and after the fifth participant, about 83 percent of the problems have been identified. This claim is backed up not only by this mathematical formula but by anecdotal evidence as well. Many usability professionals only test with five or six participants during an iterative design process. In this situation, it is relatively uncommon to test with more than a dozen, with a few exceptions. If the scope of the product is particularly large or if there are distinctly different audiences, then a strong case can be made for testing with more than five participants.

5.8.2 Five Participants Is *Not* Enough

More recently, some researchers have challenged this idea of the magic number 5 (Molich et al., 1998; Spool & Schroeder, 2001; Woolrych & Cockton, 2001). Spool and Schroeder (2001) asked participants to purchase various types of products, such as CDs and DVDs, at three different electronics websites. They discovered only 35 percent of the usability issues after the first five participants—far lower than the 80 percent predicted by Nielsen (2000). However, in this study the scope of the websites being evaluated was very large, even though the task of buying something was very well defined. Woolrych and Cockton (2001) discount the assertion that five participants is enough, primarily because it does not take into account individual differences among them.

The analyses by Lindgaard and Chattratichart (2007) of the nine usability tests from CUE-4 also raise doubts about the magic number 5. They compared the results of two teams, A and H, that both did very well, uncovering 42 and 43 percent, respectively, of the full set of usability problems. Team A used only 6 participants, whereas Team H used 12. At first glance, this might be seen as evidence for the magic number 5, since a team that tested only 6 participants uncovered as many problems as a team that tested 12. But a more detailed analysis reveals a different conclusion. In looking specifically at the overlap of usability issues between just these two reports, they found only 28 percent in common. More than 70 percent of the problems were uncovered by only one of the two teams, ruling out the possibility of the 5-participant rule applying in this case.

5.8.3 *Our* Recommendation

In our experience, five participants *per significantly different class of user* is usually enough to uncover the most important usability issues. In most of the usability tests we're conducted over the years, regardless of the total number of test participants, we're seen most of the significant issues after the first four or five participants. In fact, it is a rare occurrence when we see a new and significant issue during the fifth or sixth usability session. When we do test more than five participants for a single test, we usually see a major dropoff in attendance from the observers after about the fifth participant. Seeing the same issues over

CALCULATING *p*, OR PROBABILITY OF DETECTION

Calculating the probability of detection is fairly straightforward. Simply line up all the usability issues discovered during the test. Then, for each participant, mark how many of the issues were observed with that participant. Add the total number of issues identified with each participant, and then divide by the total number of issues. Each test participant will have encountered anywhere from 0 to 100 percent of the issues. Then, take the average for all the test participants. This is the overall probability rate for the test. Consider the example shown in this table.

Participant	Issue 1	Issue 2	Issue 3	Issue 4	Issue 5	Issue 6	Issue 7	Issue 8	Issue 9	Issue 10	Proportion
P1	x		x		x		x	x	x		0.6
P2	x	x		x		x					0.4
P3			x			x		x	x	x	0.5
P4	x	x			x			x	x	x	0.6
P5				x	x		x				0.3
P6	x					x		x	x		0.4
P7		x	x		x			x	x	x	0.6
P8	x		x	x	x		x	x		x	0.7
P9		x	x	x		x	x		x		0.6
P10		x			x						0.2
Proportion	0.5	0.5	0.5	0.4	0.6	0.4	0.4	0.6	0.6	0.4	0.49

Once the average proportion has been determined (0.49 in this case), the next step is to calculate how many users are needed to identify a certain percentage of issues. Use the following formula:

$$1 - (1 - p)^n$$

where n is the number of users.

So if you want to know the proportion of issues that would be identified by a sample of three users:

- $1 - (1 - 0.49)^3$
- $1 - (0.51)^3$
- $1 - 0.133$
- 0.867, or about 87 percent, of the issues would be identified with a sample of three users from this study

and over isn't much fun for anyone. From our very unscientific sample, we do seem to find support for the magic number 5.

The magic number 5 has worked well for us but only under the following conditions:

The scope of the evaluation is fairly limited. This means we are not doing a product-wide assessment, but rather looking only at a limited set of functions—usually about 5 to 10 tasks and about 20 to 30 web pages.

The user audience is well defined and represented. If we pretty much know who we want to test with, and they are well represented in testing, then five is adequate. If we identify more than one unique audience, then we will strive to have about five participants from each user group. Of course, the challenge is knowing when you really do have different user groups. For example, does it matter whether the user is retired? In our experience with websites, sometimes it does and sometimes it doesn't.

5.9 SUMMARY

Many usability professionals make their living by identifying usability issues and by providing actionable recommendations for improvement. Providing metrics around usability issues is not commonly done, but it can easily be incorporated into anyone's routine. Measuring usability issues helps you answer some fundamental questions about how good (or bad) the design is, how it is changing with each design iteration, and where to focus resources to remedy the outstanding problems.

You should keep the following points in mind when identifying, measuring, and presenting usability issues.

1. The easiest way to identify usability issues is during an in-person lab study, but it can also be done using verbatim comments in an automated study. The more you understand the domain, the easier it will be to spot the issues. Having multiple observers is very helpful in identifying issues.
2. When trying to figure out whether an issue is real, ask yourself whether there is a consistent story behind the user's thought process and behavior. If the story is reasonable, then the issue is likely to be real.
3. The severity of an issue can be determined in several ways. Severity always should take into account the impact on the user experience. Additional factors, such as frequency of use, impact on the business, and persistence, may also be considered. Some severity ratings are based on a simple high/medium/low rating system. Other systems are number based.
4. Some common ways to measure usability issues are measuring the frequency of unique issues, the percentage of participants who experience a specific issue, and the frequency of issues for different tasks or categories of issue. Additional analysis can be performed on high-severity issues or on how issues change from one design iteration to another.
5. When identifying usability issues, questions about consistency and bias may arise. Bias can come from many sources, and there can be a general lack of agreement on what constitutes an issue. Therefore, it's important to work collaboratively as a team, focusing on high-priority issues, and to understand how different sources of bias impact conclusions. Maximizing task coverage may be key.