# Practical Usability Rating by Experts (PURE): A Pragmatic Approach for Scoring Product Usability

**Christian P. Rohrer**
Chief Design Officer
Intel Security
Santa Clara, CA 95054, USA
christian.p.rohrer@intel.com

**Frederick Boyle**
Principal Experience Researcher
Intel Security
Santa Clara, CA 95054, USA
fritz.boyle@intel.com

**James Wendt**
User Experience Researcher
Intel Security
Santa Clara, CA 95054, USA
james.t.wendt@intel.com

**Sara Cole**
Senior Manager, User Research
Intel Security
Santa Clara, CA 95054, USA
sara.m.cole@intel.com

**Jeff Sauro**
Founding Principal
MeasuringU
Denver, CO 80206, USA
jeff@measuringusability.com


ACM 978-1-4503-4082-3/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2851581.2851607

## Abstract
Usability testing has long been considered a gold standard in evaluating the ease of use of software and websites—producing metrics to benchmark the experience and identifying areas for improvement. However, logistical complexities and costs can make frequent usability testing infeasible. Alternatives to usability testing include various forms of expert reviews that identify usability problems but fail to provide task performance metrics. This case study describes a method by which multiple teams of trained evaluators generated task usability ratings and compared them to metrics collected from an independently run usability test on three software products. Although inter-rater reliability ranged from modest to strong and the correlation between actual and predicted metrics did establish fair concurrent validity, opportunities for improved reliability and validity were identified. By establishing clear guidelines, this method can provide a useful usability rating for a range of products across multiple platforms, without costing significant time or money.

## Author Keywords
Measuring usability; benchmarking; usability rating; expert reviews

## ACM Classification Keywords

H.5.2. INFORMATION INTERFACES AND PRESENTATION (e.g., HCI) > User Interfaces > Benchmarking

## Introduction

As user experience professionals working for a large company that produces solutions on many platforms, our focus is on building a variety of software applications across form factors and operating systems. Intel Security©, like many other companies, wants to assess the quality of these applications and places high value on quantitative characterizations of product usability that can be continually improved upon over time; however, quantifying the user experience of a product has its challenges. Traditional methods of measuring user experience based on behavioral and attitudinal metrics can be difficult and costly to implement regularly in a practical setting, where the volume of new applications being produced continues to increase. Automated methods for testing large numbers of users on a website are not always applicable for benchmarking desktop and mobile applications because such tools are not always available across the range of platforms being tested. With these constraints in mind, our team set out to develop a quantitative method of scoring user experience that accurately represents the quality of a product's user experience that is not as resource-intensive as traditional methods.

To measure the user experience, we first needed to define the measurable components of user experience. Our team has been using a simple user experience model consisting of three key components: appeal, usability, and usefulness (Figure 1).

- Appeal: This can be considered the outermost layer of the experience, heavily influenced by the visual design and content of the product. Sometimes called "the look, feel, and sound" of the product, these elements contribute to its appeal and lead the user to make an almost immediate judgment of the product's quality.
- Usability: ISO 9241-11 has defined usability as the "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction" (International Organization for Standardization, 1998). We have narrowed this definition of usability slightly, describing it as the ability for the target user to be able to use the product easily.
- Usefulness: In this context, we define usefulness as the degree to which the product meets a specified user need.

Although these three components interact with each other and are essential to creating a great user experience, the focus of this case study is on our approach to scoring the usability component of user experience. We will discuss our goals for this evaluation method, our decision to use internal raters rather than testing with users, lessons learned along the way, and the impact this practical usability rating by experts (PURE) has had within our organization.

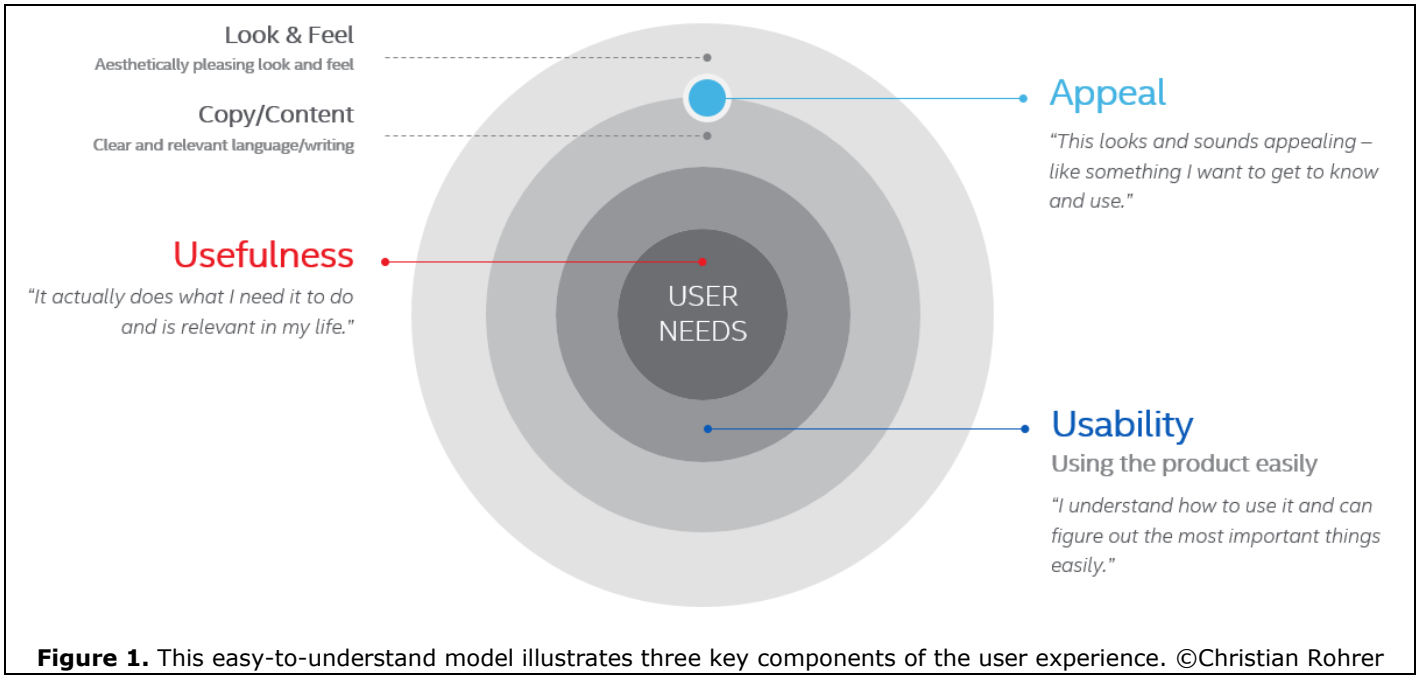


**Figure 1.** This easy-to-understand model illustrates three key components of the user experience. ©Christian Rohrer

## Goals of Implementing our Usability Rating Method

One of our major goals for this new approach was to find a way of quantitatively scoring usability that would be feasible to implement within our organization and to repeat with regularity. Traditional methods for benchmarking a product's usability, although valuable, can be expensive and time consuming. Between recruiting participants, running the testing sessions, and analyzing the results, traditional benchmarking can take weeks to complete. This becomes even more logistically challenging if one wanted to conduct such evaluations across multiple versions of a product or against competitors.

## Internal Raters Versus Testing Users

In keeping with our goal of making a lightweight method capable of being used frequently, we chose to use expert evaluators to rate each product's usability, rather than bringing in users and running them through traditional benchmark tests. Given the number of products for which our team was responsible, and the limited time and money to produce such assessments, this approach best fit our circumstances.

Our decision was informed by a rich history in the usability literature on using analytic as opposed to empirical methods to uncover problems in an interface (Hollingsed & Novick, 2007). Popular methods include

some variety of an expert reviewing the interface: heuristic evaluations (Nielsen & Molich, 1990), cognitive walkthroughs (Lewis et al., 1990), and guideline reviews (Bastien & Scapin, 1995).

In a heuristic evaluation, an expert in usability principles reviews an interface against a set of broad principles called heuristics. These heuristics are typically derived from an examination of many problems uncovered in usability tests to generate overall principles. The expert then inspects the website to determine how well it conforms to these heuristics and identifies shortcomings (Nielsen, 1993).

A cognitive walkthrough is a usability inspection method similar to a heuristic evaluation, but with the emphasis on task-scenarios that users would likely perform with the software or website (Lewis et al., 1990). Prior to conducting a cognitive walkthrough, the evaluator must first identify the users' goals and how they would attempt to accomplish the goals in the interface. An expert in usability principles then meticulously goes through each step, identifying problems users might encounter as they learn to use the interface.

A guideline review involves having an evaluator compare an interface against a detailed set of guidelines. Guidelines can be used for creating an interface (typically used by designers and developers) or evaluating it for compliance (typically performed by usability evaluators). Guideline reviews predate the web and became more popular with the increase in graphical user interfaces (GUIs). One of the best known and most comprehensive set of guidelines was sponsored by the U.S. Air Force and MITRE Corporation. Published in 1986, *Guidelines for Designing User Interface Software* contains 944 mostly usability-related guidelines (Smith & Mosier, 1986). Apple released their *Human Interface Guidelines* one year later (Apple Computer, 1987), followed by Microsoft (Microsoft Corporation, 1995).

A number of studies compare the effectiveness of each inspection method and the various problems uncovered vis-à-vis usability tests (e.g., Jeffries et al., 1991; John & Marks, 1997; Karat et al., 1992). One theme that emerged in these studies was the high variability in results among evaluators. Nielsen and Molich (1990) warned that any single evaluator is unlikely to uncover most of the usability problems. They recommended using between three and five evaluators. More recent research has found that multiple evaluators conducting heuristic evaluations independently tend to find between 30% and 50% of the problems also found in a concurrently run usability test (Law & Hvannberg, 2004; Sauro, 2012).

There is evidence that using more detailed guidelines improves the quality of inspection methods. Bastien and Scapin (1995) found that evaluators following guidelines uncovered more problems than those who just inspected the interface. They argued that ergonomic-based guidelines can act as a framework for evaluators by reducing the variability and increasing the ability to detect issues. Jeffries et al (1991) found that a guidelines-based approach forces a more careful examination of the interface relative to heuristic evaluations or cognitive walkthroughs.

## Internal Agreement on What to Rate

After our decision to use internal raters, we still needed to determine what specifically the raters would be rating. A key first step in this process was to get agreement on what mattered most from stakeholders, namely the product manager and design lead for whichever product we were seeking to evaluate. We wanted stakeholders to agree on two major questions:

1. Who is your target audience?
2. What are the 5–10 tasks that your target audience must be able to accomplish with the product for both the user *and* the business to be successful? (We call these "fundamental tasks.")

Getting this upfront agreement provided our team with some very useful constraints. First, it helped us avoid the common pitfall of stakeholders saying their product is "for everyone." Evaluating a product's usability is far easier if the rater can keep the target user in mind. Additionally, defining the 5–10 most fundamental tasks enabled us to focus the evaluation only on what was most important. This was particularly helpful in keeping the method lightweight. Perhaps of even greater value to the organization was that getting this agreement served to unite everyone in the company around a common vision for the products being built, specifically around what mattered most and for which users.

## Defining our Target Audience

To help define our target audience, we opted to select them from a set of personas recently developed by our organization. These personas were created using a combination of ethnographic field studies and quantitative market segmentation, and illustrated key user behaviors, attitudes, and motivators specific to our domain. Having these personas readily available allowed us to better frame discussions with stakeholders around who their target users were and created a common language across product teams. Describing the target audience in the form of personas resulted in raters being more effective at adopting the perspective of these users, because personas, as a means of characterizing users, are easier to empathize with than, say, demographic or purely statistical characterizations of the target audience.

## Evaluation

With our target audience and fundamental tasks identified, it was time for our evaluators to conduct their assessment. To aid in this process, we developed a rubric for objectively assessing a product's usability. Once trained on the rubric, the evaluators reviewed and rated each fundamental task for the product being evaluated.

For every fundamental task, there will be a number of "steps" required for a user to complete the task. We defined a "step" as some active decision the user has to make to keep progressing in the task. For example, creating an account is a fairly common fundamental task. Within this task, there might be several steps such as creating a username and password, agreeing to a license agreement, or granting an app permission to send push notifications. During the evaluation, our raters rated each step of a given task on a simple 1-to-3 scale (Table 1), as follows:

| Rating | Definition |
|---|---|
| 1 | Easy to understand and perform, either because the process is relatively simple and the call to action is clear, or the interaction pattern is familiar and the response is learned, such as an End User License Agreement page |
| 2 | Requires some cognitive effort to process and figure out but is doable for most users |
| 3 | Very hard to understand for most people because it does not fit an expected or typical pattern, or it has multiple calls to action |

**Table 1**. Each step is rated on a scale of 1–3.

Once each step in a task is rated, the score for each step is summed to provide on overall "task score" and color. For example, if a task consists of four steps and each step in the task was given a rating of 2, the task score would be an 8. Additionally, each step rating is given a corresponding color:

- 1 = green
- 2 = orange
- 3 = red

The color that appears most frequently in the task is considered the "dominant" color and becomes the color for the task score (Figure 2).


**Figure 2.** The task score can be interpreted somewhat like a golf score, that is, lower is better. More importantly, green is best for an overall task color; red, in contrast, is a sign the task is simply too hard for target users.

Another decision our team made was to follow what was considered the "happy path" for each task. This was done to maintain consistency each time the evaluators rated the product. If the evaluators rate a task one week and guess a particular step incorrectly four times, then rate it another week and get lucky on their first guess, the product should not necessarily get a different rating each time. The trade-off in making this decision is that not following other paths leads to a less comprehensive evaluation, but there is more consistency across what evaluators are rating. In other words, whereas following every single path would potentially uncover more potential issues, it would go against our goal of making a method lightweight enough to repeat with regularity.

During the evaluation itself, each rater had a spreadsheet in front of them to log their ratings along with any corresponding comments. One evaluator served as the "driver" and was responsible for actually interacting with the product being assessed. During the evaluation, evaluators kept their ratings private, communicating only to confirm that they were rating the same step at the same time. The team also used GoToMeeting© to record the evaluation session so each step could be referred back to and confirmed during the reporting process.

## Scorecard

Once the evaluation was completed, the ratings were compiled into a comprehensive scorecard showcasing individual task scores along with a total usability score, which summarized all fundamental task usability for the product in question (Figure 3).

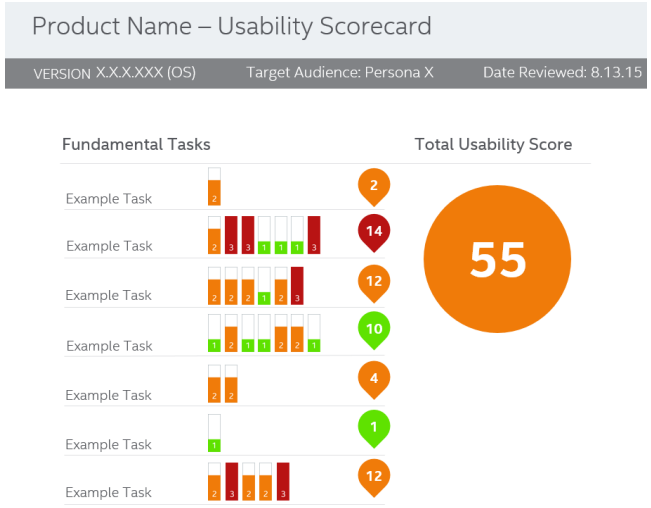


**Figure 3.** The usability scorecard shows the score for each fundamental task along with a total usability score.

As with each task, a dominant color was given to the total usability score based on which color was most prevalent among the fundamental tasks. (Ties go to the more difficult color.)

## Common Questions from Stakeholders

As we have socialized these scores among our various product teams, several common questions have come up that are deserving of an answer.

*Q: What is a good score?*

A: Simply put, it is like golf: Lower is better and green is good.

*Q: What is this number out of? 100?*

A: No, this number is influenced by how many fundamental tasks exist, how many steps exist within each task, and how difficult each step is determined to be.

*Q: What can I compare this to?*

A: The best comparison would be to compare this score against itself over time or against competitors that are evaluated. It would not make sense to compare scores across products because the number and nature of fundamental tasks are likely to differ.

## Journey Toward Increased Validity and Reliability

We have tweaked this method many times since first trying it. In one of our earliest iterations, our entire design team (30 people including designers, researchers, writers, and business analysts) performed the evaluation. In this particular attempt, we went around the room for each step and had every team member state their rating out loud. Needless to say, this yielded some wildly varied answers and little consistency.

We did notice, however, higher inter-rater reliability among the ratings given by our team's user researchers. Because this was ultimately a method for measuring usability, we opted to perform the evaluation with just our researchers several more times over the next few months, adding elements that

increased the consistency of our scores. This included not sharing ratings during the evaluation itself, clearly defining our target audience, and clarifying what constitutes a step in the broader task.

We then worked with an outside agency (MeasuringU©) for two purposes: First, we wanted to have them run a more traditional usability benchmarking study alongside our evaluation method to determine if the rating method was producing valid results. Second, we wanted to determine if outside raters could be trained on the method and reliably reproduce our results. Overall, our rubric showed good reliability. The average inter-rater reliability between four MeasuringU evaluators was $r = 0.5$ on the first attempt. (Internally, our inter-rater reliability was consistently between 0.6 and 0.88). Our rubric also showed fair concurrent validity with moderate correlations between task ease ($r = 0.48$) and overall system usability scale/standardized user experience percentile rank questionnaire (SUS/SUPR-Q) scores ($r = 0.4$). That said, the MeasuringU team approached this method a bit differently than the Intel team. For example, of the four MeasuringU evaluators, two conducted the evaluation at the same time in the same room, whereas the other two evaluators rated the product at a separate time on their own. We believe that more detailed training on the method would increase its validity and reliability.

## Impact

The most exciting outcome of using this method has been the shift in conversation among product teams and executives on where to allocate our engineering resources, now that they have a measure of usability to focus on. Providing teams with a quantitative score makes them want to improve upon that score (Figure 4). Although the underlying issues being uncovered with this method are largely the same as what we have found in traditional usability testing, the quantified nature of the findings have given them newfound visibility among the many other measures of product success.

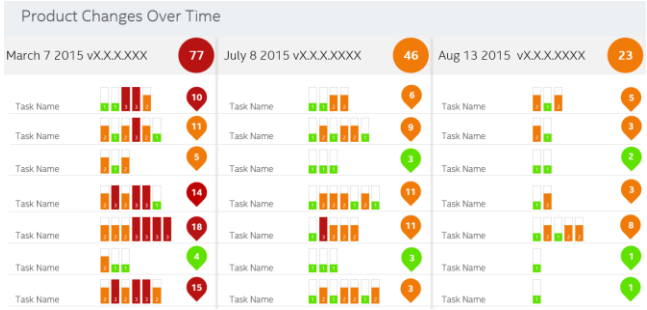


**Figure 4.** By displaying a product's usability scorecard across versions, teams can more directly see how changes they make improve their product's usability

## Acknowledgments

We thank all members of the Intel Security Consumer Experionce Design team for their support with this project. Particular thanks go to Christopher Buswell and Dustin Vaughn-Luma for their help with the visual designs used in reporting scores, as well as Alexandra Hunter and Susan Simon-Daniels for editing. Thank you also to the extended MeasuringU team for all their help.

## References

1. Apple Computer. 1987. *Human Interface Guidelines: The Apple Desktop Interface*. Addison-Wesley, Reading, MA.
2. J. M. Christian Bastien and Dominique L. Scapin. 1995. Evaluating a user interface with ergonomic criteria. *Int. J. Hum.-Comput. Interact.* 7, 2 (April 1995), 105–121. DOI:http://dx.doi.org/10.1080/10447319509526114
3. Tasha Hollingsed and David G. Novick. 2007. Usability inspection methods after 15 years of research and practice. In *Proceedings of the 25th Annual ACM International Conference on Design of Communication* (SIGDOC '07). ACM Press, New York, NY, 249–255. DOI:http://dx.doi.org/10.1145/1297144.1297200
4. International Organization for Standardization. 1998. *ISO 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs)—Part 11: Guidance on usability*. International Organization for Standardization, Genève, Switzerland. Retrieved December 2, 2015 from https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-1:v1:en
5. Robin Jeffries, James R. Miller, Cathleen Wharton, and Kathy M. Uyeda. 1991. User interface evaluation in the real world: A comparison of four techniques. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '91). ACM Press, New York, NY, 119–124. DOI:http://dx.doi.org/10.1145/108844.108862
6. B. E. John and S. J. Marks. 1997. Tracking the effectiveness of usability evaluation methods. *Behaviour and Information Technology*, 16, 4/5, 188–202.
7. Claire-Marie Karat, Robert Campbell, and Tarra Fiegel. 1992. Comparison of empirical testing and walkthrough methods in user interface evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '92). ACM Press, New York, NY, 397–404. DOI:http://dx.doi.org/0.1145/142750.142873
8. Effie L.-C. Law and Ebba T. Hvannberg. 2004. Analysis of strategies for improving and estimating the effectiveness of heuristic evaluation. In *Proceedings of the Third Nordic Conference on Human-Computer Interaction* (NordiCHI '04). ACM Press, New York, NY, 241–250. DOI:http://dx.doi.org/10.1145/1028014.1028051
9. Clayton Lewis, Peter G. Polson, Cathleen Wharton, and John Rieman. 1990. Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '90). ACM Press, New York, NY, 235–242. DOI:http://dx.doi.org/10.1145/97243.97279
10. Microsoft Corporation. 1995. *The Windows Interface: Guidelines for Software Design*. Microsoft Press, Redmond, WA.
11. Jakob Nielsen. 1993. *Usability Engineering*. Academic Press, Boston, MA.
12. Jakob Nielsen and Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in*

*Computing Systems* (CHI '90). ACM Press, New York, NY, 249–256. DOI:http://dx.doi.org/10.1145/97243.97281

13. Jeff Sauro. 2012. How Effective are Heuristic Evaluations? (September 2012). Retrieved August 10, 2015 from http://www.measuringu.com/blog/effective-he.php
14. Sydney L. Smith and Jane N. Mosier. 1986. *Guidelines for Designing User Interface Software*. MITRE Technical Report MTR-10090. The MITRE Corporation, Bedford, MA.