

Quantifying and Comparing Ease of Use Without Breaking the Bank

by Christian P. Rohrer, PhD

Summary: The PURE method quantifies how difficult a product is to use and provides qualitative insights into how to fix it, both without costing a lot of time or money.

The Business Context

Face it: Businesses need metrics in order to operate. When it comes to a company's user experience, the desire to measure is just as strong, despite the [risks of doing so](#). As a result, the use of [analytics](#), [A/B testing](#), surveys, and [usability metrics](#) have all grown significantly over the years. This practice is likely to persist, if not grow further, which makes it worthwhile to scrutinize the metrics we use and consider what's missing to meet the goal of truly measuring user experience.

Broadly speaking, [traditional metrics](#) can be broken down into behavioral (what people do) or attitudinal (what people say) measures. Behavioral metrics are gathered from usage, as users perform actions on software or websites, and are commonly used in analytics and A/B testing. They include counts (users, page views, visits, downloads), rates ([bounces](#), conversion, installation, task success), and times (time on page, time on task, engagement). Common attitudinal measures come from surveys ([Net Promoter Score](#), System Usability Scale, customer [satisfaction](#)) or user ratings. While these are all useful, there are significant limitations:

1. Numbers alone don't usually provide the insights needed to understand *why* an effect was observed or how to fix a problem.
2. The metrics used in analytics and A/B testing are typically *indirect* indicators of the quality of the user experience: they reflect software performance, not human experience.
3. Classic measures of user experience, such as those derived from usability benchmarking studies, are expensive and time-consuming, so they aren't used frequently enough to provide regular assessment and tracking.

PURE (Pragmatic Usability Rating by Experts) is a relatively new usability-evaluation method that attempts to sidestep these problems in a way that is reasonably quick, cheap, reliable, and valid. The metrics resulted from PURE can be used frequently and comparatively, making it practical to publish metrics for each version of a product or across a set of competitors, with just a few days of effort. When used with other measures, PURE scores fill in an important gap left by the limitations of traditional metrics.

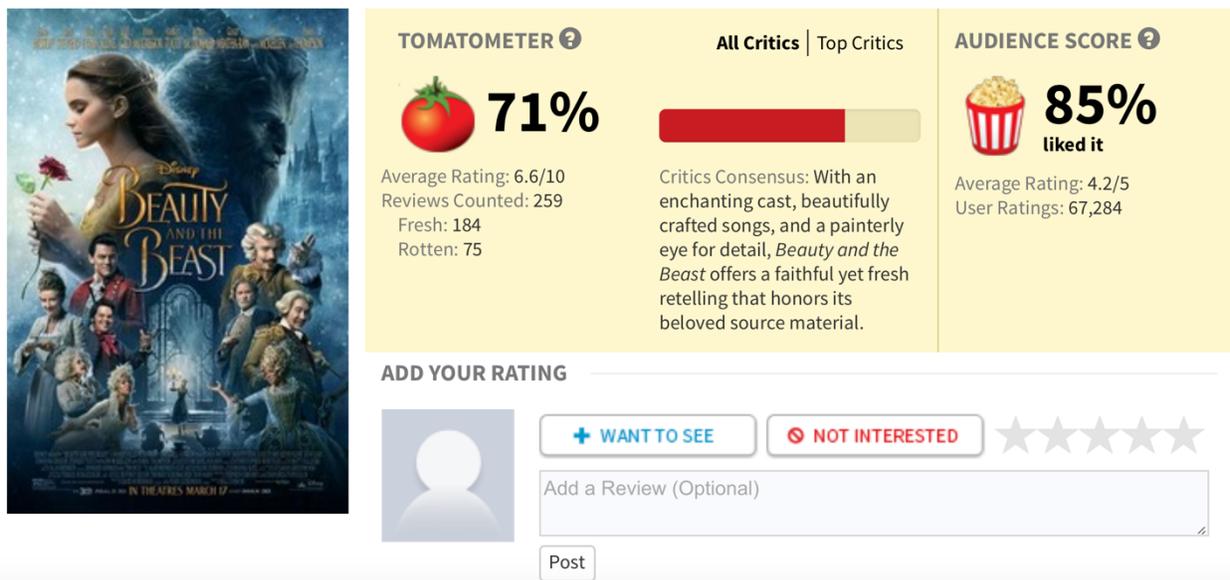
The PURE Method: Metrics of Another Kind

Attitudinal and behavioral metrics are not the only way to produce useful numbers to represent user-experience quality. Another type of metric can serve a similar purpose, but is much more

practical to generate: one based on a type of expert review, resulting in a detailed rating of an experience. This is the basis for the PURE method.

Definition: PURE is a usability-evaluation method in which usability experts assign one or more quantitative ratings to a design based on a set of criteria and then combine all these ratings into a final score and easy-to-understand visual representation.

To understand PURE, consider an analogy in the movie-going experience. Movies are judged by popularity and by how they are perceived by both critics and audiences. In particular, to quantify critic appeal, the movie-review site Rotten Tomatoes features a *Tomatometer*, indicating the percentage of approved movie critics who have given the movie a positive review. The site also includes an *Audience Score*, showing the percentage of users who liked the movie.



The screenshot displays the Rotten Tomatoes interface for the movie *Beauty and the Beast*. On the left is the movie poster. The main content area is divided into three columns. The first column, titled 'TOMATOMETER', shows a 71% rating with a tomato icon, an average rating of 6.6/10, 259 reviews (184 fresh, 75 rotten), and a bar chart. The second column, 'All Critics | Top Critics', contains a 'Critics Consensus' paragraph. The third column, 'AUDIENCE SCORE', shows an 85% rating with a popcorn icon, an average rating of 4.2/5, and 67,284 user ratings. Below these sections is an 'ADD YOUR RATING' section with a user profile icon, buttons for '+ WANT TO SEE' and 'NOT INTERESTED', a five-star rating system, a text input field for 'Add a Review (Optional)', and a 'Post' button.

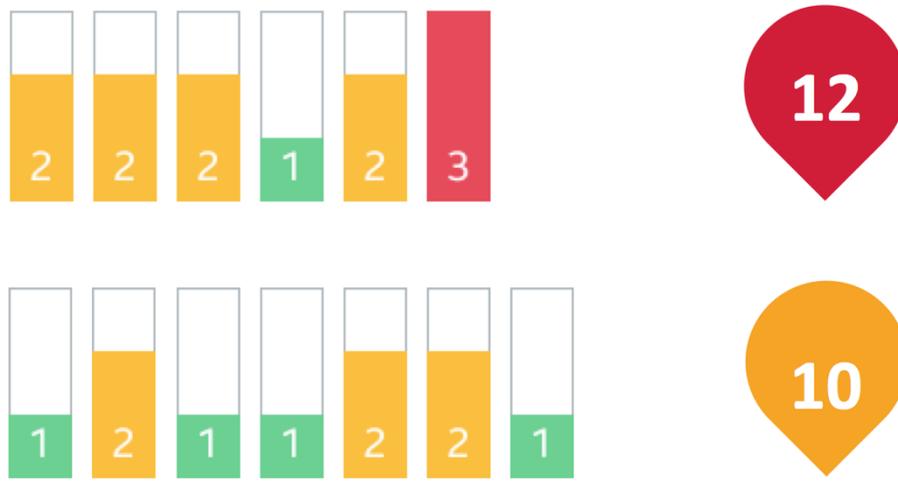
Rotten Tomatoes assigns each movie two metrics: the Tomatometer, reflecting critics' perception, and the Audience Score, representing audience ratings.

PURE is like a *Tomatometer* for usability. It provides a score obtained by aggregating ratings (on a predefined rating scale) from a panel of experts familiar with UX principles and [heuristics](#). Similarly, when rating a movie, critics consider elements common in all movies — such as plot, acting, entertainment value, aesthetics, technical aspects, and social relevance.

One important difference, however, is that, unlike movie critics who rate movies based on their own preferences and world views, in the PURE method expert raters attempt to provide a score representing how good the experience would be for a specific, well-defined target user type. This approach increases the consistency and reliability of the ratings provided by a PURE panel reviewing the same experience and also allows the PURE scores to be legitimately used for comparison purposes.

PURE Scores: Measures of Friction

PURE is focused on just one component of user experience: ease of use. Other aspects of user experience, such as aesthetic appeal, effectiveness (meeting user needs), or resulting emotions are not addressed. But having a measure of ease of use is critical, because, if the target users aren't able to easily use a given product or service, they cannot unlock its potential benefits. Here's an example of PURE scores for two tasks supported by a product or service:



PURE scores for two tasks

Each task has a series of colored bars, which represent a step in that particular task. Each of those steps is rated and colored, based on how easy or difficult that step is judged to be for the target user. The rating on each step is based on a simple 1–3 scale, defined by the following scoring rubric:



The step can be **accomplished easily** by the target user, due to low [cognitive load](#) or because it's a known pattern, such as the acceptance of a terms-of-service agreement.



The step requires a **notable degree of cognitive load** (or physical effort) by the target user, but can generally be accomplished with some effort.

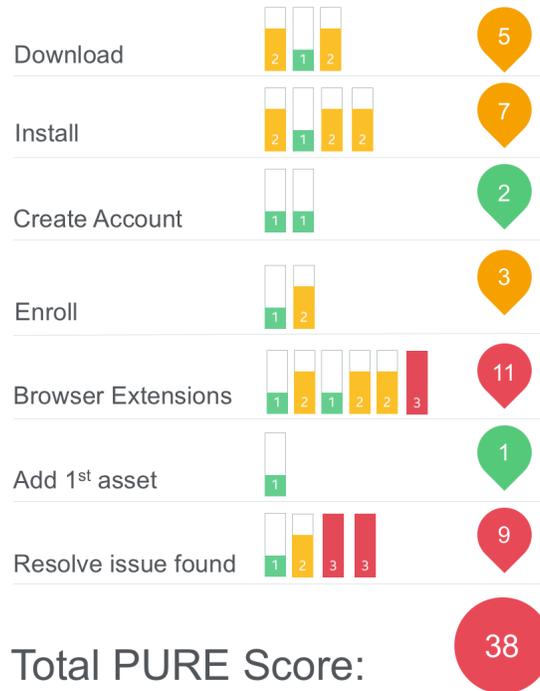
3

The step is **difficult for the target user**, due to significant cognitive load or confusion; some target users would likely fail or abandon the task at this point.

The **PURE score for a given task is simply the sum of the scores** of all steps' ratings in that task. The **color** of the task is determined by the **worst rating** score in the task. For example, a single step rated a red 3 causes the whole task to take on the red color.

The numbers and colors shown in PURE scores represent *friction*, the opposite of ease of use. The higher the number and the “hotter” the colors, the more friction there is — similar to [usability-severity ratings](#). [Comparing the PURE scorecard for the same task across different product versions or among competitors allows you to easily see the variation in friction for different designs of the task. Although lower numbers usually mean less friction, the quality of the steps should also be considered, as indicated by their colors. One of the big benefits of PURE is that it considers overall user effort, rather than just clicks or steps. This can help counter overly simplistic arguments that fewer clicks will result in higher levels of success, and instead refocus attention on reducing *user effort*, rather than just clicks. \(Note that you should generally avoid comparing the PURE scores of different tasks, since their nature and goals are often quite different. \)](#)

Because PURE measures the friction in a set of tasks, it is important to define the tasks to be reviewed. Pragmatically, not every task can be measured, so in PURE, we only score the “fundamental tasks” — those critical for the target user and the business. Here is a sample PURE score for a product with 7 fundamental tasks:



The PURE score for a product is the sum of the scores for each fundamental task that can be accomplished with that product.

The PURE score for the product (38 in this case) is the sum of the PURE scores for all fundamental tasks. Just like for tasks, the overall color for the product is determined by the worst color of the fundamental tasks in the product. This means that a single red step (rated 3) in any fundamental task causes that entire task *and* product to be colored red. The rationale for this convention is that no consumer product should have a step in which the target user is likely to fail a fundamental task. The color red has a tendency to make that statement clearly and focus attention to potential points of failure in the product.

If this sounds difficult to understand or explain, consider a simple analogy: PURE is like golf — lower numbers are better, and green is good.

The PURE Method's Impact on Business Practices

When business stakeholders have an easy to understand, numeric representation of an important aspect of their product or service, like ease of use, they tend to be highly motivated to improve on it, and may set goals to do so. This response is the same for *any* metric — whether PURE scores or other traditional metrics, like the total number of users or minutes used per week. Stakeholders will want to improve these numbers as well. But, unlike these other metrics, PURE scores are *operational* — they show what caused poor metrics and where the user experience needs improvement, providing a clear roadmap for refining the design. Showing PURE at regular business meetings, where product or business metrics are discussed, helps ensure that projects aimed at improving user experience are prioritized and executed.

Once the PURE method is learned, it is relatively easy to conduct a PURE evaluation and to compare PURE scores on competing products, because competitors typically have the same set of fundamental tasks and the same target audience. Business stakeholders are even more motivated to address issues identified by PURE when they see how their product stacks up against the competition and what they need to improve to win. The competitive nature of business culture becomes a significant ally in the effort to build great user experiences.

Another benefit of PURE is that you can use it on user experiences that haven't been completely built yet. While it is more accurate when conducted on fully functioning products, PURE can be applied to [medium-fidelity prototypes](#) or to clickable wireframes — to either compare possible solutions to the same design problem or see how a proposed flow fares in terms of ease of use before committing to coding it.

How to Conduct a PURE Evaluation

Using the PURE method to score a given product or service requires certain steps to be taken, many of which are helpful for any crossfunctional product, design, and development team. There are 8 required and 2 optional steps to follow:

1. Clearly identify the target user type(s).
2. Select the fundamental tasks of this product for target users.
3. Indicate the happy path (or the desired path) for each fundamental task.
4. Determine step boundaries for each task and label them in a PURE scoresheet.
5. Collect PURE scores from three expert raters who walk through the happy paths of the fundamental tasks together and silently rate each step.
6. Calculate the interrater reliability for the raters' independent scores to ensure reasonable agreement among experts.
7. Have the the expert panel discuss ratings and rationale for individual scores, and then agree on a single score for each step.
8. Sum the PURE scores for each fundamental task and for the entire product; color each step, task, and product appropriately.
9. (Optional) For each step, provide a screenshot (or photo) and a qualitative summary of the experts' rationale for the scoring of that step.
10. (Optional) If comparing multiple products or product versions, prepare a comparative PURE scorecard, showing the same PURE task scores side by side.

Here is a little more detail on each of these steps.

Step 1: Target User Types

In order for the PURE experts to consistently estimate ease of use, they must have a specific user type in mind. Assumed user qualities such as technological savvy or familiarity with the current product will heavily impact the experts' evaluations. For example, users who have forgotten their passwords may be asked to enter a one-time passcode texted to their

smartphones. Those familiar with this pattern will find it relatively easy, but users never exposed to it may encounter difficulties. The target user type can be supplied by the product manager or the lead designer, or can be decided upon by the expert team. The decision and any assumptions about the target user's context must be documented, especially if future comparative PURE scores are expected. Target user types based on [personas](#) work well, because they are usually easier to understand and already familiar to the team. However, a well-defined user description can work in practice as well, as long as behaviors and rationale for those behaviors can be understood by the expert panel. A clear target user type will help the PURE raters be consistent in their ratings, and it has the additional benefit of getting crossfunctional teams to agree on who is most important for their product or service and why. It is possible to identify and score against multiple target user types. This doesn't necessarily double or triple the work involved in PURE, but it will significantly increase it. In practice, target user types should be limited to no more than 2–3, or this method loses its "pragmatic" nature.

Step 2: Fundamental Tasks

Fundamental tasks are defined as tasks that either:

- are critical for the business to succeed (e.g., payment/checkout), or
- allow target users to meet their core needs (assuming the product or service offers a value proposition that meets some of those needs).

For most consumer products and mobile apps, there are typically less than 10 fundamental tasks, and they form the basis on which a PURE score is generated. However, websites or complex applications may have more than 10 fundamental tasks. My recommendation is to keep the number of tasks close to 10, and no more than 20, at least when you first start using this method. It's always possible to add more fundamental tasks, if your first attempt at PURE is found valuable (though you shouldn't try to compare product PURE scores without explaining that new tasks were added to later analyses).

Step 3: Happy Paths

A given task can often be fulfilled in a variety of ways, with a different number of steps for each method. PURE requires the team to identify the "happy path," which is the most desired way in which the target user would accomplish this task. This path is our best shot at making the task easy for users, so it makes sense to focus PURE scoring on this particular flow more than on any other.

It would be reasonable to evaluate multiple paths for the same task, but, just like having more than one target user type, doing so increases the time and effort required to conduct a PURE evaluation. Also, other methods, like heuristic evaluation or standard usability studies, would be sufficient to find and fix problems in other paths. I would only use PURE on multiple paths if it seemed critical to measure and compare them.

Lastly, some teams have chosen to use PURE to evaluate the "popular path" by looking at clickstream analytics to determine which flow is most likely for a given goal. This is a reasonable decision, and it may take the place of a happy path for some teams.

Step 4: Step Boundaries

Once the happy paths are determined, it is critical to go through them and identify where each step begins and ends. Depending on the type of interaction provided by the product or service, this process can be harder than you might think. The place to start is the “default step” definition:

- A **step begins** when a system presents the user with a set of options (e.g., a user interface is rendered).
- A **step ends** when the user takes an action and expects significant system response to that action.
- A step may contain microinteractions, such as manipulating form fields; these are considered part of the step.

This definition works for a majority of tasks on screen-based interfaces, but may require some refinement for certain situations. That’s OK, as long as you document why you deviated from the default definition of the step, so you can repeat this decision in later PURE analyses. Also, there can sometimes be debate about what “expected significant system response” is. For example, a web page may hide and show content as users interact with certain elements, which can be unexpected. Or, when a long page contains several sections, it is tempting to call each section a step. However, keep in mind that changing the page sectioning may affect future PURE scores, so be clear on the implications of choosing those step boundaries. The lead researcher should make the decision and document it for future PURE-scoring consistency.

Step 5: Review by Three Expert Raters

The PURE method uses three usability experts (ideally UX researchers) to provide the initial PURE scores. It is important that the designers or other product professionals responsible for the rated flow NOT be on the panel of experts, because, in practice, it is very hard to be objective about one’s own designs.

The panel assembles (either in the same room, or remotely) and all members watch as the lead researcher goes through each task on a common screen and declares when each step begins and ends. Each panel member silently rates and reviews each step, making notes about the rationale for the rating. The notes can include observed usability problems, as in [heuristic evaluations](#).

One big difference from heuristic evaluation is that, in PURE, the panel sees the same experience together, which ensures that they rate the same thing — otherwise their scores would be wildly different. This point highlights an important difference in goals between heuristic evaluation and PURE. In heuristic evaluation, the goal is to find as many usability issues as possible and get a complete view of the usability of the product or service. In contrast, PURE aims to provide a reliable measure of how easy it is for the most important user type to accomplish only the fundamental tasks, through the best design offered at this point. The analysis starts here, because these are most important areas to get right. Once improved, the PURE method can be used elsewhere, although it may not be necessary to provide a numerical score for all paths, tasks, and user types, once buy-in to address ease of use is achieved in general.

Experts should be able to enter their PURE scores without being exposed to other experts' scores. This is easily accomplished with the use of an online spreadsheet, with tabs for each rater. Task names are propagated to each tab, and all scores can be automatically rendered onto a master tab for review in later steps.

Step 6: Interrater Reliability Calculations

To see how much the experts agreed with each other in their individual PURE scores, which were provided silently, you should calculate the “interrater reliability” (IRR). IRR is a measure of how much the raters agree, given their understanding of the target user type and the 1–3 rubric. While this calculation may seem overly academic, it does ensure that there is a reasonable level of agreement among experts, and is important for methodological soundness. Reviewing this number will help the experts understand whether they made the same assumptions as they rated products in PURE.

There is more than one way to calculate IRR. I recommend using Krippendorff's alpha. To compute it, you can use a free online calculator such as [ReCal](#) (select “ordinal” data type, since the 1–3 rubric is ordinal).

IRR ranges from -1 to 1, but is typically between 0.5 and 1.0. If the experts are not able to achieve an IRR of at least 0.667, they should discuss why they varied so much and simply consider this PURE evaluation to be a training session. It typically takes 2–3 rounds of trying PURE before a panel of experts has sufficiently understood the rubric and the user type to be consistent, so plan for a few rounds of trial and error for learning purposes.

Step 7: The Decided PURE Score

After the experts have recorded their individual ratings, they should walk through the steps of each task and discuss them together. This discussion is invaluable for two main reasons: (1) expert raters will learn from their colleagues, and, over time, will become better and more consistent raters; and (2) the expert panel will be able to decide on a single score for each step. This “decided score” will be the reported PURE score for this step, and it will benefit from the collective wisdom of the entire panel of expert raters.

The decided score is easiest to determine when all raters gave the same individual score. If 2 out of 3 agreed on a score, it usually becomes the decided score, but not always. Sometimes the ensuing discussion may cause the team to choose the less popular score as the decided score. This situation most often happens when a specific assumption or key insight about the experience was missed by 2 raters, but is explained by the other rater.

Very infrequently, all three raters will have three different scores. This is almost always due to a different understanding of the method or set of assumptions, which the discussion will no doubt clarify. As with other aspects of PURE, the assumptions decided upon should be documented for future review and PURE scoring.

The PURE method does not use the average rating from the three expert reviewers for some good reasons — some pertain to the ordinal nature of the 1–3 rating scale, but, most importantly, an average would take away from the power of the decided PURE score, which represents the collective wisdom of three expert reviewers, rather than their average, undiscussed assessment.

Step 8: Summing It Up into Green, Yellow, and Red

The next step is simple and gratifying: summing up the decided PURE step scores into the task PURE scores, and then summing the task PURE scores into the product PURE score. Because these numbers are not normalized, they can be as large as the tasks and their complexity are. There is almost always room for improvement, once these numbers are summed and shown. Just as important is the visual representation of the PURE scores. Using bar heights and colors to represent friction has shown to be very effective at conveying trouble spots in a given product. Red is particularly troublesome when you consider that PURE is focused only on the most important user types, and their potential experience on the most important tasks with the company's best shot at accomplishing these tasks.

If many steps are rated a yellow 2, it means the target user had to spend some degree of effort to get through them. This may be OK or even unavoidable, depending on the nature of the task, but it will be important for the team to really consider what it can do to improve the ease of use. Even steps that are rated a green 1 can have areas of improvement. It is wise to ensure that the notes from the expert team are collated and used in the next part of the PURE method.

Step 9 (optional): It's Not All Quantitative

As raters go through each step of each task from the perspective of the target user, they should capture their rationale and notes for their scores and call out areas that could be improved in the user experience. These observations describe exactly what could be addressed to improve the PURE score, and ostensibly, the usability of the product. The expert panel should collect a screenshot or photo of each step, and document these observations into the appendix of the PURE report to help design and development teams know why and how to make improvements. Ideally, PURE scores are generated by a panel of usability experts who have seen the product perform in qualitative usability studies. It is not a requirement to have witnessed the specific product in other studies, but, at a minimum, PURE reviewers should have a deep understanding of user experience and usability, and be well versed in design [principles](#) and general [heuristics](#).

An expert reviewer who has been exposed to user studies (such as standard usability testing) on the same product will be able to apply insights from such studies to the PURE evaluation of the product. Used together, qualitative usability studies and a separate PURE analysis of the same or similar tasks can complement each other and provide in-depth information about the main usability hurdles in a design. This combination of methods can be cost-effective and time-saving, especially compared to traditional quantitative approaches, such as usability benchmarking.

Step 10 (optional): Comparing PURE Scores

One of the most gratifying aspects of the PURE method is comparing scores of the same task among product versions or competitive products, especially when improvement on one's own product is demonstrated. Below is an example of one of the first PURE scorecards to be conducted, on an actual product that went to market, after showing drastic improvements in ease of use over 5 months. The task names were genericized for confidentiality reasons, but you can see that big improvements were realized through redesign iterations by simplifying some of the task flows (cutting steps) and also improving ease of use for individual steps.

March 7 2015 v0.8.5.289	77	July 8 2015 v0.9.1.357	53	Aug 13 2015 v1.0.2.007	25
Download/Install					
Initial enrollment					
Add first entry					
Import database					
Install companion sw					
Upgrade to premium					
Resolve issue X					

Scores from three PURE evaluations on the same product indicate significant UX improvements in the new versions of the product.

Is PURE Valid and Reliable?

While the metrics defined and described here are not as precise as empirical measures based on user data, they are directionally accurate, and have been shown to have reasonable validity and reliability scores. When comparing PURE results with metrics obtained from running a usability-benchmarking study on the same product, we found statistically significant correlations with SEQ and SUS (popular ease-of-use survey measures) of 0.5 ($p < 0.05$) and 0.4 ($p < 0.01$), respectively. These numbers show that PURE has at least reasonable validity, when compared with standard quantitative metrics, at statistically significant levels ($p < 0.05$). Interrater reliability calculations for PURE have ranged from 0.5 to 0.9, and are generally very high (above 0.8), after expert raters are trained on the method. The PURE method was first documented in a [case study](#) that I published at CHI 2016 together with my coauthors James Wendt, Jeff Sauro, Frederick Boyle, and Sara Cole. In a recent PURE evaluation at Capital One, three experts achieved an interrater reliability score of 1.0 (100% agreement) across 9 fundamental tasks. As of this writing, PURE is known to have been used with over 15 different products at 3 companies. I expect to see this number grow, as the practice becomes better understood and is improved by new adopters.

Conclusion

While learning to conduct the PURE method takes some effort and not everyone is qualified to do it, experience has shown it to be an extremely valuable tool to use along with the [landscape of user-research methods](#). PURE scores capitalize on the ever present appetite for quantitative metrics and provide concrete numbers that orient the organization toward fixing ease-of-use barriers. In the end, everyone benefits: users, employees, and business stakeholders.

Sometimes all it takes is showing the right metrics in an easy to understand format, with just enough frequency to effect significant positive change.