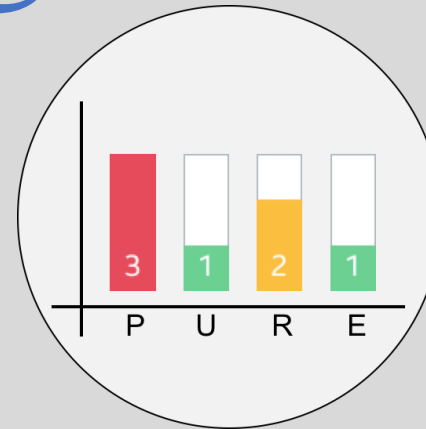


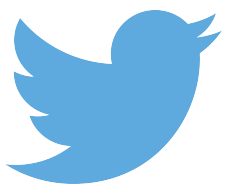
The PURE Method

Version 1.5, last edited by Christian Rohrer on 10/5/2016

PRAGMATIC
USABILITY
RATINGS by
EXPERTS



A Practical
Approach
for Scoring
Product
Usability



Continue the conversation online on Twitter and Google Groups:

#puremethod @christianrohrer @J_Wendt33 @MeasuringU

<http://www.xdstrategy.com/PURE>

<https://groups.google.com/d/forum/pure-method>



I'd like to start by saying thanks...

There were many great contributions to help bring the PURE Method to life. I would like to thank them and the companies they work at for making this possible (and others I missed here):

Michelle Bayles-Simon, Mike Benjamin, Yaro Brock, Emily Grace, Ashley Henry, Mave Houston, Samuel Jaffee, Helen Kim, David Lessard, Pallavi Kutty, Hilary Masland, Clifton McDaniel-Neff, Lucy Oh, Aaron Rigg, Susanna Rogers, Nicole Sharratt, Emily Short, Lauren Singer, Liz Thomas, Cassi Tramm, Dustin Vaughn-Luma, Orralyn Vithyavuthi, Vanessa Whatley, Lorie Whitaker, Cindy Zolnowski



Christian P. Rohrer | crohrer@yahoo.com | @christianrohrer

Jimmy Wendt | james.t.wendt@intel.com | @J_Wendt33

Jeff Sauro | jeff@measuringu.com | @MeasuringU

Frederick Boyle | Sara Cole

(Original authors of the PURE Method Case Study at CHI2016)

- And thank YOU for coming today!
- Notice a potential improvement?
- Help us make PURE better:
 1. Write down your thoughts on a sticky note
 2. Place it in your booklet on the appropriate page (or on top, if a general comment)
 3. During a break or at the end, transfer your sticky note to the same page of the master booklet in the front of the room



Today's Agenda

- Introduction and Rationale for The PURE Method
- Overview of the PURE Method
- Small Group Exercise: Applying PURE to a real product
- Calculating and improving Inter-rater reliability and establishing standards for reporting PURE Scores
- Typical PURE “set up” challenges and how to handle them
- Business Impact and Future Advances for the PURE Method
- Large Group Discussion and Closing Remarks

Business leaders, product professionals and engineers are obsessed with dashboards, metrics, quantitative tests and scores.

- “You can’t **manage**, what you can’t **measure**.”
- “I need a **dashboard** to manage and control my business.”
- “How does our **NPS** compare?”
- “Invest in **data scientists** and **Big Data**.”
- “I want to be more **scientific**.”
- “Build, **Measure**, Learn.”
- “This design needs to be **validated**.”



Definitions of “User Experience” are comprehensive, and they do not themselves easily to measurement

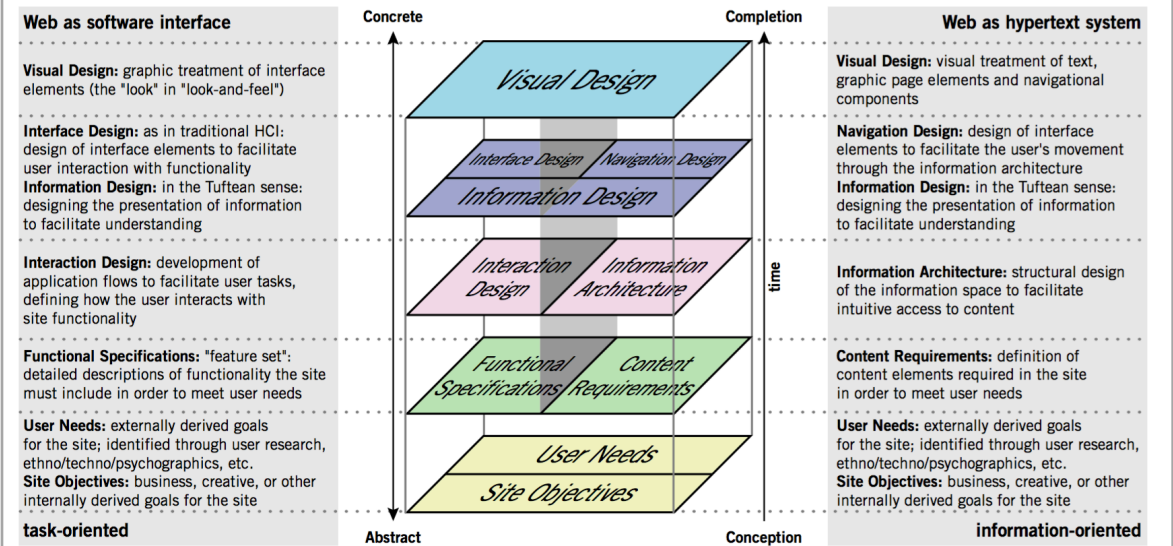
- “**All the aspects** of **how people use** an interactive product: the way it feels in their hands, how well they understand how it works, how they feel about it while they’re using it, how well it serves their purposes, and how well it fits into the entire context in which they are using it.” –Alben
- “**All aspects** of the end-user’s **interaction** with the company, its services, and its products.” –Nielsen-Norman Group
- “The **overall experience**, in general or specifics, a user, customer, or audience member has with a product, service, or event.” –Shedroff
- “**Every aspect of the user’s interaction** with a product, service, or company that make up the user’s perceptions of the whole” –UPA
- “The **overall perception** and **comprehensive interaction** an individual has with a company, service or product.” –Goto
- “Encompasses **all aspects** of a digital product that users experience directly—and perceive, learn, and use—including its form, behavior, and content. Learnability, usability, usefulness, and aesthetic appeal are key factors in users’ experience of a product.” –UXMatters
- “The **value** derived from interaction(s) with a product or service and the supporting cast in the context of use (e.g., time, location, and user disposition). –Sward & MacArthur”

Models of User Experience

The Elements of User Experience Jesse James Garrett
jig@jig.net

A basic duality: The Web was originally conceived as a hypertextual information space; but the development of increasingly sophisticated front- and back-end technologies has fostered its use as a remote software interface. This dual nature has led to much confusion, as user experience practitioners have attempted to adapt their terminology to cases beyond the scope of its original application. The goal of this document is to define some of these terms within their appropriate contexts, and to clarify the underlying relationships among these various elements.

Jesse James Garrett
jig@jig.net
30 March 2000



This picture is incomplete: The model outlined here does not account for secondary considerations (such as those arising during technical or content development) that may influence decisions during user experience development. Also, this model does not describe a development process, nor does it define roles within a user experience development team. Rather, it seeks to define the key considerations that go into the development of user experience on the Web today.

© 2000 Jesse James Garrett

<http://www.jig.net/ia/>

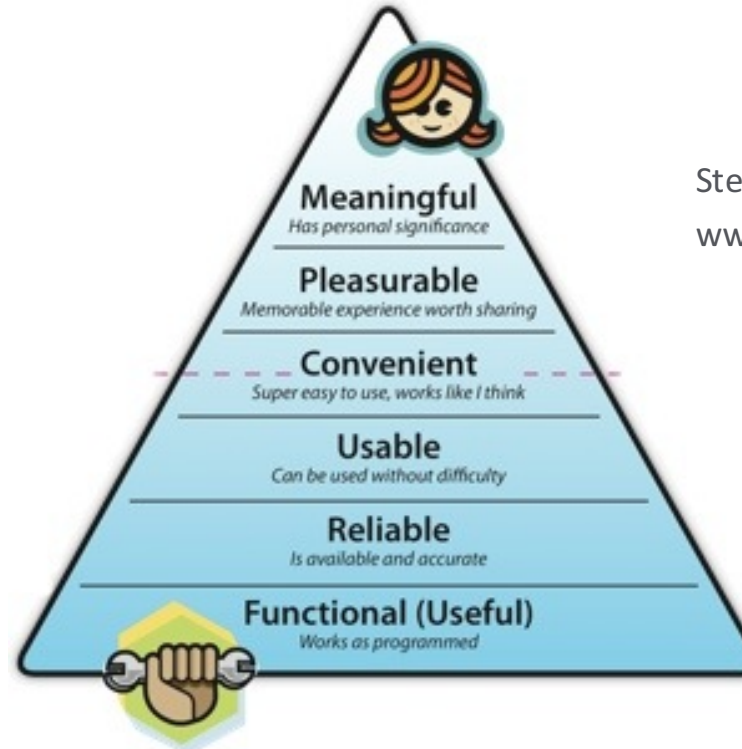
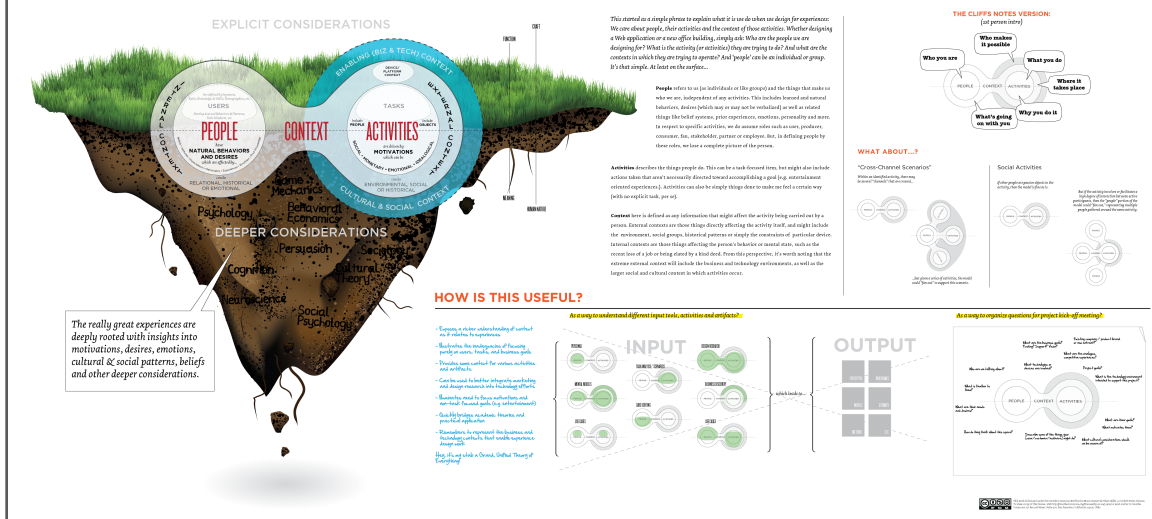
Jesse James Garrett's Elements of User Experience

2000

THE FUNDAMENTALS *of* EXPERIENCE DESIGN

by Stephen P. Anderson

"Designing for experiences is fundamentally about people, their activities, and the context of those activities..."



Stephen P. Anderson
www.poetpainter.com

Models of User Experience (cont.)



Effectiveness



Ease

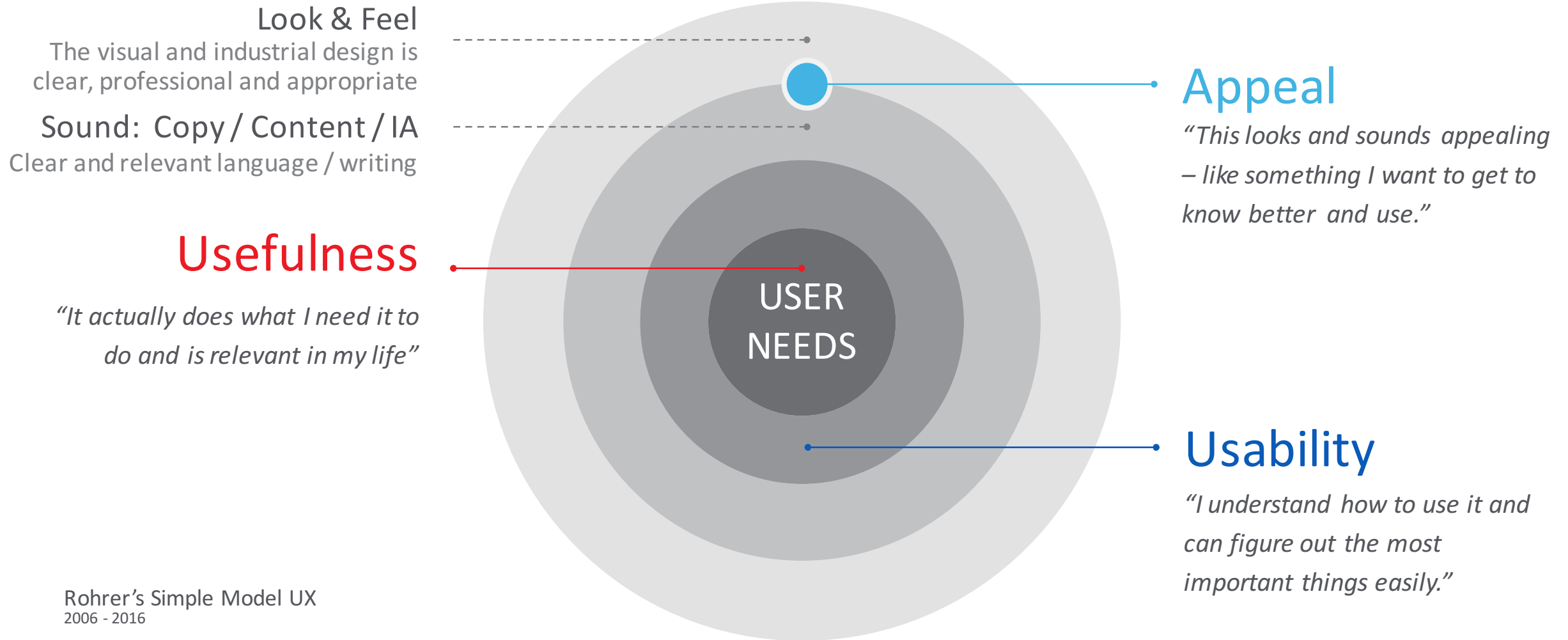


Emotion

forrester.com/cxindex

FORRESTER®

I use a Model of User Experience that is easy to explain and whose components are reasonable to measure/rate



Every one of these models includes
Ease Of Use (one of the core
concepts in usability)

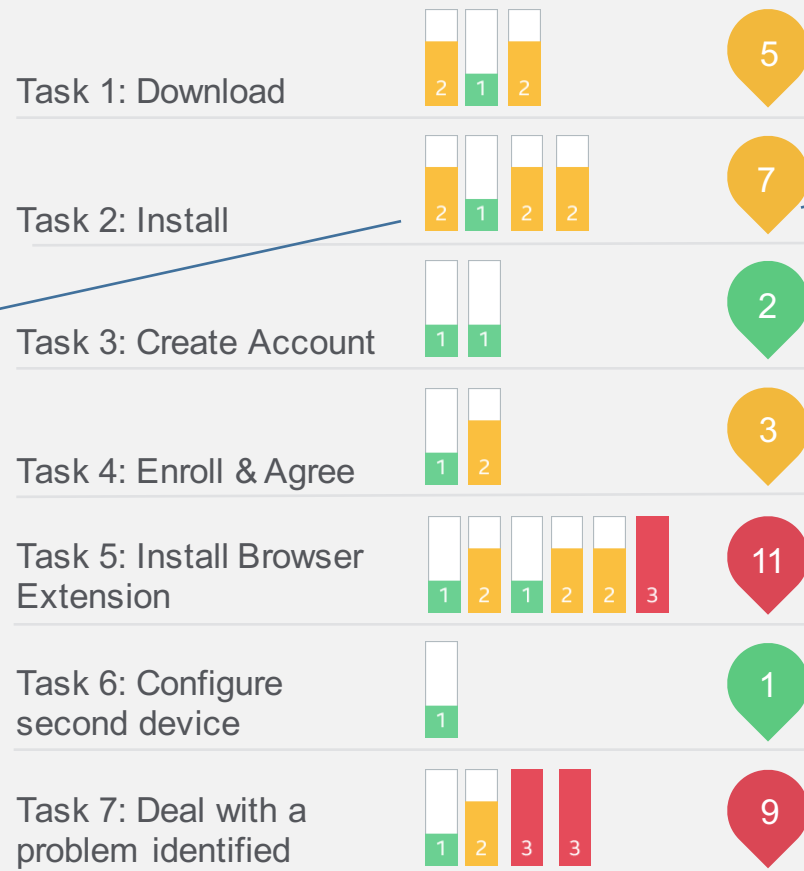
PURE is focused on the opposite of
“Ease”: Friction or Cognitive Load

PURE in brief: A quickly produced ease of use scorecard for the best performance of a product's fundamental tasks for a given user type

Only the most fundamental tasks are scored

Both the length (number of bars) and ease of use (numbers/colors) of the tasks are quickly discerned

Version, target user type and date are documented for comparison later



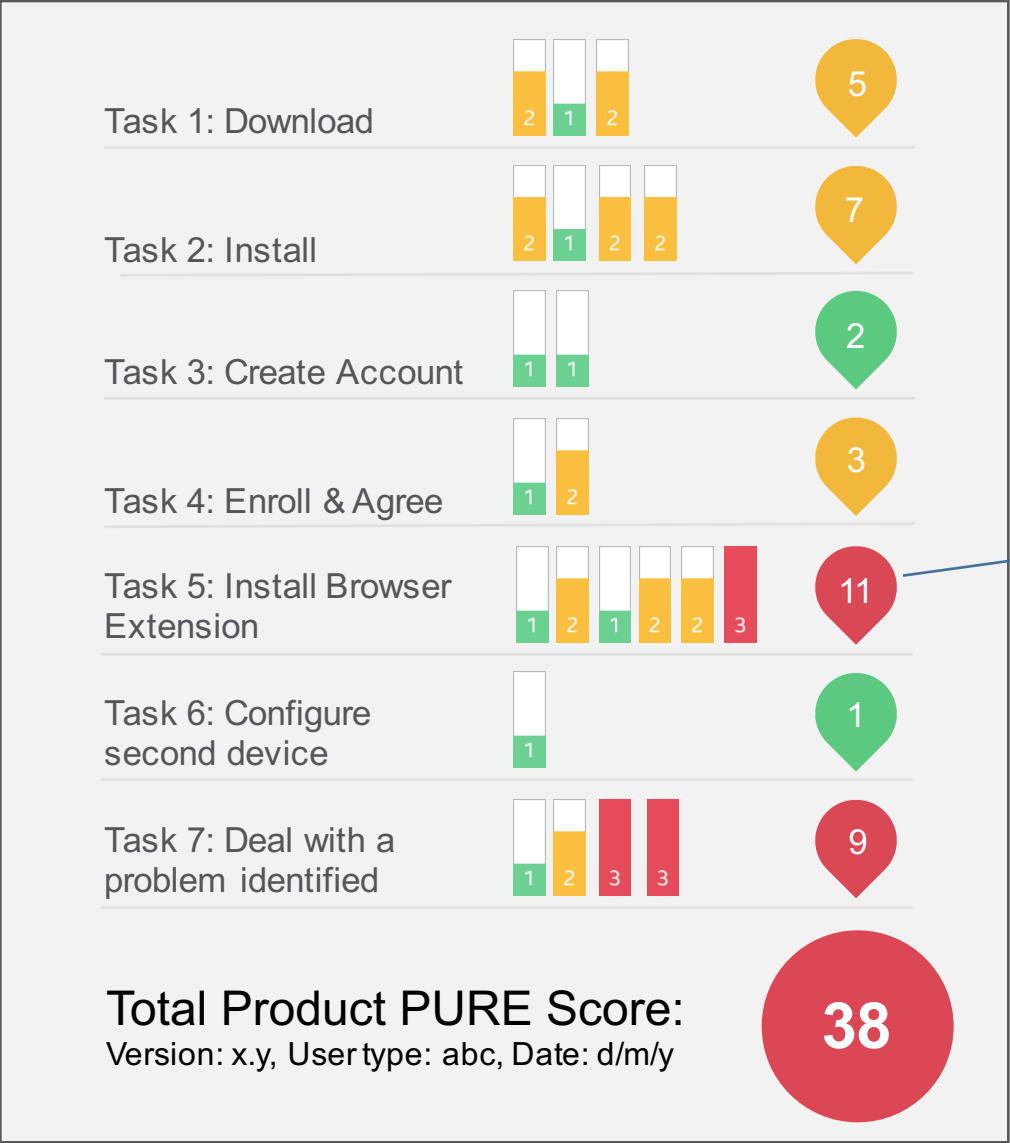
Total Product PURE Score:
Version: x.y, User type: abc, Date: d/m/y

38

Each task gets a score and color. As in golf, green is good, smaller numbers are better

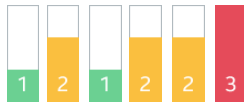
Behind every rating and score, there are helpful reasons for the scores, which drive fixes

The whole product PURE score is driven by the task PURE scores



Behind every rating and score, there are helpful reasons for the scores, which drive fixes

Task 5: Install Browser Extension



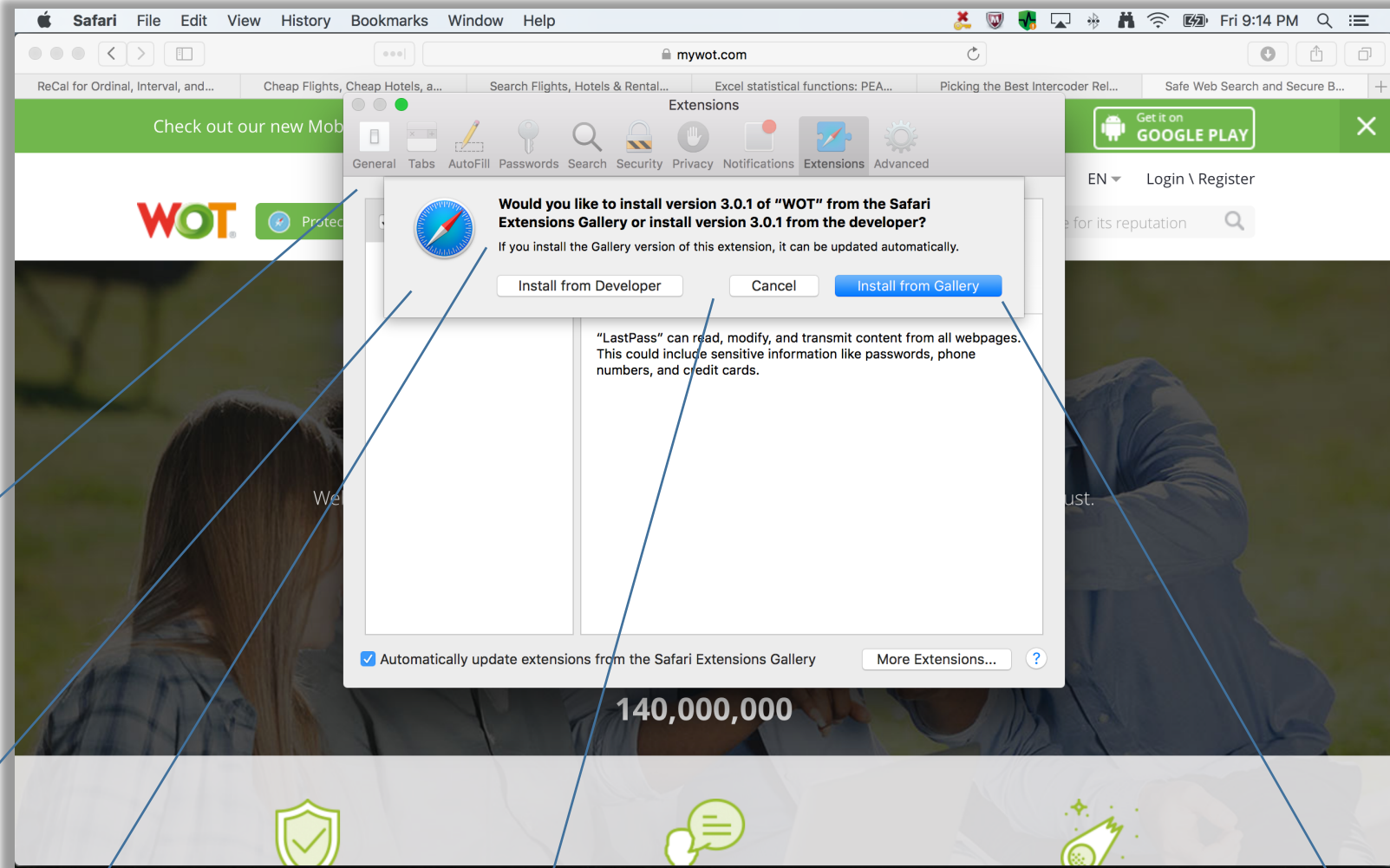
11

Why was this step a 3?

This Extensions tab from Safari settings comes out of nowhere after the previous step (unexpected)

The dialogue box appears at the same time as the Safari settings tab, partially obscuring its content and masking the context the dialog is related to.

The language used here is difficult for the target user to fully understand without significant effort. (At least a benefit is explained, however.)



There are three choices, not uniformly spaced (so looks sloppy). Most problematic, it's likely not clear to this user type what "Cancel" does at this point without some cognitive effort.

After selecting the default button "Install from Gallery" both the dialogue and the Safari setting disappear and it appears not to have done anything (this issue is technically part of the next step)

TRADITIONAL MEASURES OF USER EXPERIENCE & USABILITY

Two main approaches: Empirical Testing & Analytic Evaluation

Empirical



USABILITY BENCHMARKING



CARD SORTING



TREE TESTING



FIRST CLICK
TESTING

Analytic



HEURISTIC
EVALUATIONS



KEYSTROKE
LEVEL MODELING



PURE

Usability Benchmarks: Task-based Attitudinal and Behavioral Metrics

Attitudinal Measures



Study Level Sat.
Task Level Sat.

Behavioral Measures



Completion Rates
Errors



Time on task

Task-based Attitudinal Measure: System Usability Scale (1986)

I think that I would like to use this system frequently

I found the system unnecessarily complex

I thought the system was easy to use

I think that I would need the support of a technical person to be able to use this system

I found the various functions in this system were well integrated

I thought there was too much inconsistency in this system

I would imagine that most people would learn to use this system very quickly

I found the system very cumbersome to use

I felt very confident using the system

I needed to learn a lot of things before I could get going with this system

Strongly Disagree

1

2

3

4

5

Strongly Agree

Task-based Attitudinal Measure: SEQ



SEQ: Single Ease Question

Overall, how difficult or easy did you find this task?

Very Difficult
1

2

3

4

5

6

Very Easy
7



Task-based Attitudinal Measure: SUPR-Q

Standardized User Experience Percentile Rank Questionnaire

USABILITY

- This website is easy to use
- It easy to navigate with this website

LOYALTY

- How likely are you to recommend this website
- I will likely visit the website in the future

CREDIBILITY

- The information on this website is credible.
- The information on this website is trustworthy.

APPEARANCE

- I found the website to be attractive
- The website has a clean and simple presentation



The Empirical Methods require time, money and many users

Empirical



USABILITY BENCHMARKING



CARD SORTING



TREE TESTING



ONLINE
TESTING

Benchmark Example:

- Three product competitive benchmark (desktop SW)
- 7 tasks/scenarios
- 24 user per product
- 2+ months, \$100K

Online Testing Example:

- 1200 website users (3 sites)
- 8 tasks/scenarios
- 3 week setup and collection
- \$90K license or \$25-50K x 1

Metrics are also produced by Analytic Methods

Heuristic Evaluation

- Metrics are twofold:
 - Number of problems found by any rater
 - Severity rating (0-4) of problems

Keystroke Modeling (GOMS/KLM)

- Metrics are about time to complete tasks (based cognitive and motor)

PURE

- Scores of ease of use (friction or cognitive load) of key tasks and product

Analytic



HEURISTIC
EVALUATIONS



KEYSTROKE
LEVEL MODELING



PURE

How the Analytic Methods produce their metrics differs

Heuristic Evaluation

- 3-5 raters independently walk through an interface, to review whether good principles (heuristics) are present
- Goal is to find and document as many usability issues as possible and assign severity ratings of found problems



Keystroke Modeling (GOMS/KLM)

- 1 rater catalogs operations in an interface, using known cognitive/motor skill time limits to estimate efficiency



PURE

- 3 raters represent the perspective of a specific user type and reliably score core tasks of a product in terms of ease of use (aka friction or cognitive load)



ANALOGY: In PURE, we judge a **specific** performance, as in skating & gymnastics

- A panel of judges each silently rates a specific performance they are all witnessing
- A known rubric defines how much of a deduction results from a given mistake
- PURE rates every step, as if it were a "move"



Simone Biles' floor exercise score by panel of judges at the Rio Olympics in 2016



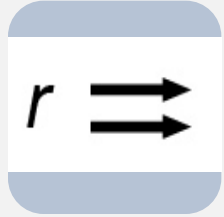
Difficulty: How difficult the moves performed in her routine, based on a panel of 2 judges. Calculated prior to the routine

+

Execution: How well she performed those move. Starting from 10 possible points, judges deduct points for each mistake

= **SCORE**

PURE Method Reliability & Validity (courtesy of MeasuringU)



Convergent Validity

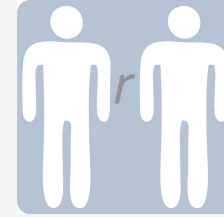
(comparing results from PURE to results from a traditional benchmarking study)

220 Users 3 Products + 8 Websites

SEQ: $r = .5$

SUS/SUPRQ $r = .4$

Completion/Time: $r < .2$



Inter-rater Reliability

(calculated both among raters inside the same company and across raters at the company and at the agency MeasuringU)

$r = .5 \text{ to } .9$

Validity: A classic usability benchmarking study was conducted by MeasuringU and values for attitudinal measures (SEQ, SUS, SUPRQ) and behavioral measures (completion time) were gathered. Then, a PURE score was conducted by MeasuringU researchers, and the results were compared. **Reliability:** MeasuringU researcher PURE scores and Intel researcher PURE scores were compared.

OVERVIEW OF THE “PURE” METHOD

Preparing to conduct the PURE Method

What you'll need:

- Collaboration from Product, Design and Technology for the Kickoff/Setup meeting and supporting materials
- At least 3 expert evaluators (user researchers, typically)
- Displays, Devices, Recording Equipment, Spreadsheets
- ~8 hours (~3-4 for the evaluation and ~4 for the report)

PURE Method Overview

1. Define the **Target User Type(s)**
2. Determine the **Fundamental Tasks**
3. Specify the “**Happy Path**” of each Fundamental Task
4. Raters **walk through** each Step of Happy Paths of all tasks and **rate** on 1-3 scale using the **PURE Rubric**
5. **Discuss** ratings, check for **inter-rater reliability**, and try another round if minimum IRR (0.667) not achieved
6. **Sum and color the ratings** for all Fundamental Tasks and the sum up the Total PURE Score for the product

Define User Type

- The Target User type is clearly defined and described by the Design Lead and Product Management (e.g., via well-known personas):
- Consider key behavioral and attitudinal attributes that are relevant to how the product will be used.



"Show me the price! I'll shop around until I get the best deal."

Tracy
Thrifty Bargain Hunter

DEMOGRAPHICS

Age	35
Family Status	Married with two children
Occupation	Receptionist
Household Income	\$60,000
City of Residence	Dayton, OH
Housing	Owns a house

Key Behaviors

- Sees price first
- Seeks bargains
- Buys items she doesn't need, but "may someday"
- Cannot pay a retail price (she's "better than that")

Identify Fundamental tasks

- The Fundamental Tasks are identified by PM and the Design Lead
- These are defined as the **5-10 tasks that the Target User(s) MUST be able to do for both the user and the business to be successful.**
- **Example Fundamental Tasks for software that monitors security of multiple devices**
 1. Discover & Learn
 2. Download & Install
 3. Create Account
 4. Add a device
 5. Access Dashboard
 6. Send Message

Specify the “Happy Path” of each Task

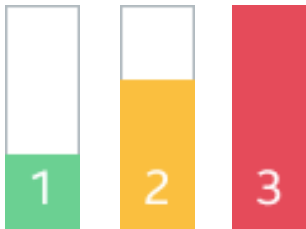
- The “**Happy Path**” of each Fundamental Task is defined as **the most desired path to accomplish the task, as specified by the Design Lead.** (Alternatively, you can choose the “typical path,” if you have analytic data that shows this. Either way, pick one path and stick with it for the analysis.)
- You may need help from Technology to invoke the Happy Path realistically (e.g., usernames/passwords or QA server access)
- We do not evaluate divergent paths, because it creates too much variation and we want to score the “best current performance.”

Example Happy Path with 6 steps:



Walk through each Step and Rate

- The three raters walk through the Happy Path of each Task
- They silently provide a 1-3 difficulty rating for each step of each Fundamental Task, based on the **PURE Rating Rubric** (next slide).
- PURE Scores are stored in a spreadsheet; inter-rater reliability is checked later.
- Each step's rating is colored green (for 1), yellow (for 2) or red (for 3).



	A	B	C	D
1	Step #	Description	Rater 1	Rater 1 Comments
2	1	Google Search for product name	1	Standard google search results page
3	2	Arrive on Product Landing Page	2	Multiple calls to action. Product comparison chart lacks discoverability
4	3	Assess product comparison chart	1	Clear breakdown of product features. Information is easily digestible
5	4	Click on "Buy Now" for Product X	1	"Buy Now" is the obvious call to action
6	5	Fill out page one of the cart	2	While creating a password, there is a lack of guidance as to the password requirements
7	6	Complete page 2 of the cart	1	Standard billing form
8	7	Confirm order	1	Order information appears correct, one clear call to action
9	8	Download product	1	Receipt page had one clear download button
10	9	Install product	1	Installer provides clear visibility into the status of the installation
11	10	"Finish" installation	3	Product fails to launch automatically after installation, leaving the user stranded
12	11	Product home page	1	Clear distinction between "Log in" and "Sign up"
13	12	Create Account	2	While creating a password, there is a lack of guidance as to the password requirements. Unclear if this account is related to the one on the website that was created earlier.
14	13			
15	14			
16	15			
17	16			

PURE Rating Rubric

A rating of 1, 2 or 3 is given for each step in the task, based on the PURE Rating Rubric:



1 = The step can be accomplished easily and quickly by the Target User, because there is very little cognitive load. Example attributes:

- Easy to understand language/user interface
- A single, recognizable and clear call to action
- A familiar interaction pattern, such as the acceptance of a EULA (end user license agreement)



2 = The step requires some degree of thought by the Target User, but can generally be accomplished with such effort



3 = The step is difficult for for the Target User, due to significant cognitive load; some of the Target Users would likely fail this task

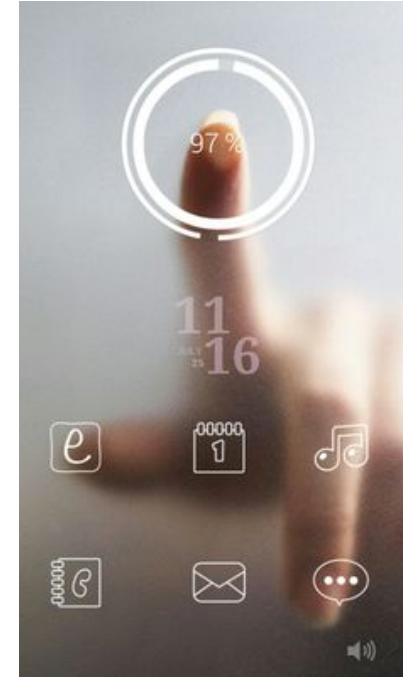
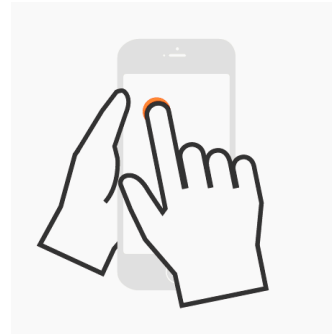
What is a “Step” in the task? The **Default Definition of Step**

A step **begins** when the system presents the user with a set of options and is waiting for user input to proceed.

STEP BEGINS:
UI is rendered



Micro-interactions may occur within the UI until the user reaches a point where they provide input to proceed to the next step



STEP ENDS:
When user makes selection and expects a significant system response

A step **ends** when the user makes a selection on the options provided AND receives an “**expected significant system response**” to that user action.

Note: micro-interactions that make it easier to accomplish parts of the step do not typically count as a step, unless you decide they do and choose to consistently score this way.

Determine a single PURE rating for each step by using the mode; if all 3 raters are different for a given step, discuss each of your rationales to solidify a common set of criteria; try again – your ratings will most likely converge.

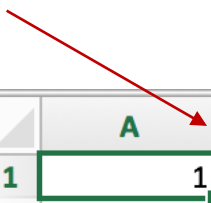
Example ratings for Task 1:

Task # & Step #	Rater 1	Rater 2	Rater 3	Avg	Mode	Decided Rating
T1 S1	1	1	1	1	1	1
T1 S2	1	1	1	1	1	1
T1 S3	3	2	3	2.66666667	3	3
T1 S4	1	1	1	1	1	1
T1 S5	2	2	1	1.66666667	2	2
T1 S6	1	1	1	1	1	1

The “mode” (i.e., the most common rating) should be the rating entered for the decided upon PURE ratings. Using the average suggests a level of precision that we don’t really have and makes it harder to represent the step rating with a simple 1, 2 or 3 (and the associated bar height & color).

Calculate Inter-rater reliability (if < 0.667 , don't report it; try again)

- Put all ratings of all tasks into a single sheet w/ no headers



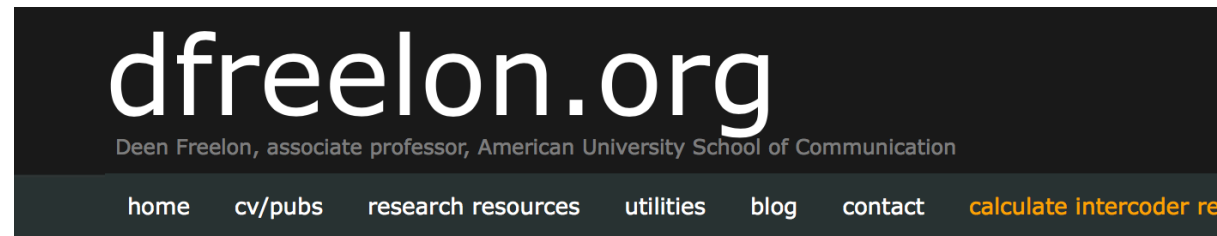
	A	B	C
1	1	1	1
2	1	1	1
3	3	2	3
4	1	1	1
5	2	2	1
6	1	1	1

- Save as a CSV

- Upload CSV file to ReCal OIR calculator:

<http://dfreelon.org/utis/recalfront/recal-oir/>

- Select “Ordinal” before uploading




ReCal for Ordinal, Interval, and Ratio Data (OIR)

ReCal OIR (“**R**eliability **C**alculator for **O**rdinal, **I**nterval, and **R**atio data”) is an online utility that computes intercoder/interrater reliability coefficients for ordinal, interval, and ratio data judged by **two or more coders**. (If you need to calculate reliability for nominal data judged by two coders only, use [ReCal2](#); for nominal data coded by three or more coders, use [ReCal3](#).) Here is a brief feature list:

- Calculates three reliability coefficients:
 - Krippendorff’s alpha for ordinal data
 - Krippendorff’s alpha for interval data
 - Krippendorff’s alpha for ratio data
- Accepts any range of possible variable values, including decimal values and negative numbers
- Results should be valid for **ordinal, interval, or ratio data sets coded by two or more coders** (other uses are not endorsed, and accurate results are not guaranteed in any case — trust but verify!)

If you have used ReCal OIR before, you may submit your data file for calculation via the form below. If you are a first-time user, please read [the documentation](#) first. (*Note: failure to format data files properly may produce incorrect results!*) You should also read ReCal’s [very short license agreement](#) before use.



☒ Ordinal ☐ Interval ☐ Ratio

no file selected

Example output from ReCal OIR (with inter-rater reliability at 0.787 – well above 0.667!)

ReCal for Ordinal, Interval, and Ratio-Level Data results for file "Interrater_Reliability_ProductX_Ver2.1.csv"

File size: 208 bytes

N coders: 3

N cases: 30

N decisions: 90

Krippendorff's alpha (ordinal) 0.787

Select another CSV file for reliability calculation below:

☒ Ordinal ☐ Interval ☐ Ratio

Choose File no file selected

Calculate Reliability

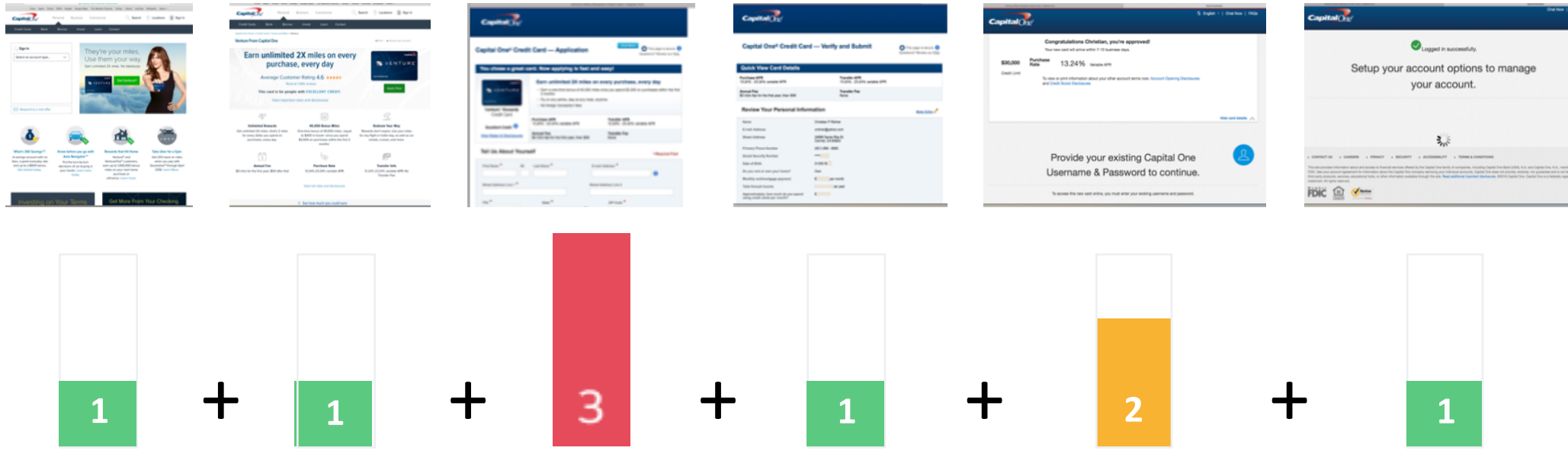
☐ Save results history ([what's this?](#))

Disclaimer: This application is provided for educational purposes only. Its author assumes no responsibility for the accuracy of the results above. You are advised to verify all reliability figures with an independent authority (e.g. a calculator) before incorporating them into any publication or presentation. If you have any questions, comments, or suggestions regarding ReCal, please send them to deen@dfreelon.org.

If you found ReCal useful, please consider [leaving a comment](#). Any and all feedback is appreciated.

PURE Scores by Task

The sum of all ratings for a given Fundamental Task is the PURE Score for that task:



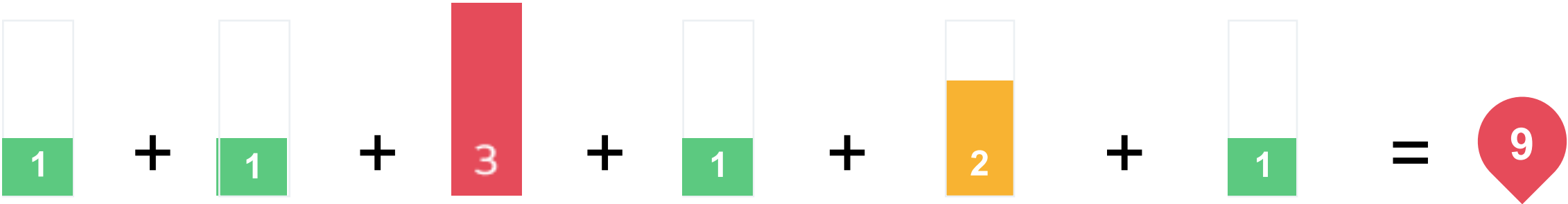
PURE Scores are assigned a color, based on the worst color of the components. One red or yellow score makes the whole task red or yellow. Like golf, smaller numbers are better and green is good.

Why do we need inter-rater reliability?

- This does two things:
 - Helps us learn how to rate more consistently with each other
 - Provides more methodological soundness to the method
- The premise of any “rubric” is that it is objective enough to be applied consistently by trained raters; used widely in education & psychology
- If our IRR is < 0.667 (using Krippendorff’s alpha), it stands on shakier ground. We need a sufficient N ($\sim 25+$), so include at least 2-3 tasks and all their steps before calculating it
- Once the rating team is trained (takes 2-3 times), they will be able to produce reliable scores time after time without a lot of effort

Sum Task PURE Scores for Total Product PURE Score

Task
1



Task
2



Task
3



Total Product PURE Score = 19

Let's try rating a task or two together

OPTION 1

- Target User: Tracy the Thrifty Bargain Hunter (see slide 30)
 - Savvy shopper, both online and offline – wherever the deals can be had
 - Knows how to use coupons and promo codes
 - Member of Groupon, LivingSocial, Nexttag, etc.
- Fundamental Task 1: Use Google to find cheapest place to buy a new pair of Asics Kayano 21 Women's, size 6.5 [skip]
- Fundamental Task 2: Find cheap turquoise or red Asics Kayano 21 Women's size 6.5, and put into cart, using Amazon.com

OPTION 2

- Target User: Savvy Traveler Steph
 - Regular domestic traveler for business (travels cross country about once a month)
 - Frequent flyer member of 4 airlines (American, United, Virgin, Southwest)
 - Uses aggregator sites (Kayak, Hipmunk) to get best deal and routing
- Fundamental Task 1: Find a “good” flight from SFO to Richmond, VA:
 - SFO to Richmond, VA (RIC)
 - Leave Mon during the day, not too early in the morning
 - Return on Fri, arriving back by 7pm
 - No redeye flights; no more than 1 stop; no long layovers
 - Use one of 4 favorite airlines, if possible
- Fundamental Task 1A: Use **Kayak**
- Fundamental Task 1B: Use **Hipmunk**
- Fundamental Task 1C: Use **Expedia**

SMALL GROUP EXERCISE (~1 hour)

- Now we will form into small groups of 3 (or 6) and attempt to perform a limited version of the PURE method
- Each table/group will have a “Lead” – see Lead Duties on next slide
- We will have a separate handout describing the User Type, Tasks, Happy Path of Tasks, and the rationale for these decisions
- The group will clarify any questions they have with each other, and the Lead will document new questions and assumptions as they arise
- Step through at least 2 tasks, calculate inter-rater reliability and be prepared to share IRR from both tasks with large group
- Comments for improvement? Go onto sticky notes; consider sharing with large group later on

Documents are here: <https://goo.gl/Qz9Qru>

NEW! Use this: <https://goo.gl/Lbd3pm>

LEAD Duties

- Document assumptions and decisions the team has made on:
 - Target User type
 - Fundamental Task Choices
 - Happy Path specification and choices
 - Why the Target User type is following this happy path
- Run the session so all raters can see the experience reasonably well
- Let raters know when a task is beginning and ending
- Let raters know when a step is beginning and ending
- Facilitate discussion afterwards
 - Counterbalance who shares their ratings first, last, etc.
 - Record and later report inter-rater reliability score

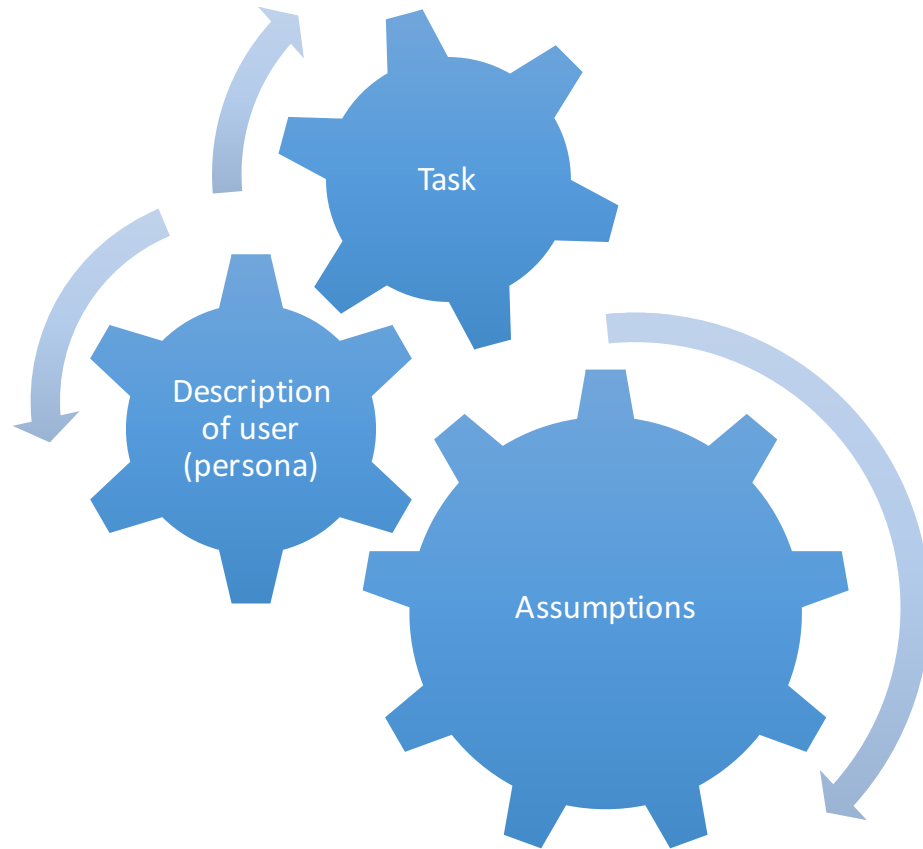
NEW! Use this: <https://goo.gl/Lbd3pm>

TYPICAL SETUP CHALLENGES

MIKE BENJAMIN

Proper set up and framing is important for an efficient evaluation

Teasing out the assumptions prior to the evaluation ensures stakeholders are aligned around the most common tasks/flows and persona, leading to a more efficient evaluation process down the line. Being more concrete and explicit about the context, the scenario, path, and steps, eliminates the guess-work during the evaluation to keep things moving properly.



Task: Make a mobile check deposit using the Capital One Mobile Banking App

Persona: User who primarily manages their account online. Occasionally uses their smartphone app to check transactions on the phone and pay bills, but has never deposited a check using mobile deposit

Assumptions: Very Important for setting up the context and narrow down the scenario

Sample kickoff discussions to tease out assumptions about our task and persona



Evaluator: I want to make sure we clearly call out the experience and the happy path we believe our user takes so we can properly set up the session



Product Manager: Keep in mind that not everybody sees the same experience, it depends on the account type and the eligibility requirements of the user. Different account types, such as Small Business, may have a different flow.



Evaluator: Great, let's concentrate on a 'consumer persona'.



Product Manager: We also need to think about users eligibility to make a deposit. That is driven by compliance, legal, and fraud, so it depends on the user and what they are trying to do at a given point. So what bucket should we look at?



Evaluator: let's assume that the user doesn't fall in any risk-buckets that prevent him from making a deposit. We'll concentrate on users who wish to deposit and have no specific restrictions



Product Manager: What about users who try to deposit via a tablet, or Android phone, the UX for Android is slightly different.



Designer: Also, we should focus on the first time experience. Those folks will see the coaching and help screens, and have to agree to the terms and conditions first time and we're aggressively marketing to get new folks in the door.



Evaluator: Great, if we can all agree, we'll narrow down the folks to reign it in



Evaluator: We narrowed it down to Mobile app users with an iPhone who are depositing a check for the first time into their personal checking account.



Designer: Do we want them to read through the all the onboarding screens and tool tips since that's part of the experience?



IT Lead: Let's narrow it to a user who skips the onboarding screens, data tells us that only 20% spend time looking at those screens. We'll factor in the Terms and conditions screens since that's a requirement for first time users.



Designer: There are 3 different ways you can get to Mobile deposit, from the mobile landing page, from the overflow menu and from the accounts page.

Sample kickoff discussions to tease out assumptions about our task and persona



Evaluator: Based on analytics, which in the most commonly used? We'll pick the one most used. Ok, so we've narrowed it to entry from within the Accounts Page, since that's the most common entry point.



Product Manager: Since you have to be logged in, are you also going to start from the 'Sign In' experience? There are several ways users can sign in in and depending on that experience the number of steps will vary



Evaluator: Great question, we're going to assume the user is an active Capital One customer, and knows their UN/PW.



IT Lead: In order to replicate the experience, you'll need the credentials to the test account:



IT Lead: You won't be able to actually submit the check and get final confirmation, because the back-end system for the UAT site, is currently on lock-down, so you can't actually submit, will that create a problem?



Evaluator: Got it. Can you provide a clickable prototype or a set of screenshots that simulates what happens after a user actually submits the check? We will be able to do it, but note that during the evaluation



Evaluator: Going through the experience, I notice that when I click the 'Mobile Deposit' button, it asks for access to my camera and GPS. When I go to allow it, it takes me through the phone's setting experience. Do we want to capture this experience as well, or assume the user has access to GPS and camera activated in settings?

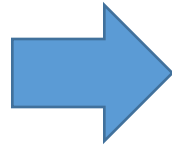


Product Manager: Let's assume user previously allowed access to GPS and has camera turned on. Data tells us that most of our folks have that setting turned on when they download our app

Proper set up and framing is important for an efficient evaluation

Task: Make a mobile check deposit using the Capital One Mobile Banking App

Persona: User who primarily manages their account online. Occasionally uses their smartphone app to check transactions on the phone and pay bills, but has never deposited a check using mobile deposit



Assumptions we teased out in the conversation:

- Evaluating first time user experience for a personal account user who's never deposited a check with any FI
- User skips initial "coaching" screens
- Using iPhone 6S w/GPS and camera app access already 'turned on'
- Persona is using UN/PW to log in, their username is stored
- Persona is not bound by account restrictions
- Persona is entering experience via 'Account' page, but we'll start at HP
- We are not able to 'submit' at this point, we will evaluate the 'Verification' and 'Confirmation' pages using screenshots and will not be able to really interact

BUSINESS IMPACT and FUTURE
IMPROVEMENTS to PURE

Leaders and Teams **want** to get better over time. An actual example:

March 7 2015 vX.X.X.XXX		77	July 8 2015 vX.X.X.XXXX		46	Aug 13 2015 vX.X.X.XXXX		23
Task Name	<div><div></div><div>1</div><div>1</div><div>3</div><div>3</div><div>2</div></div>	10	Task Name	<div><div></div><div>1</div><div>1</div><div>2</div><div>2</div></div>	6	Task Name	<div><div>2</div><div></div><div>1</div><div>2</div></div>	5
Task Name	<div><div>2</div><div></div><div>1</div><div>2</div><div>3</div><div>2</div><div>1</div></div>	11	Task Name	<div><div></div><div>1</div><div>2</div><div></div><div>1</div><div>2</div><div>2</div><div>1</div></div>	9	Task Name	<div><div>2</div><div></div><div>1</div></div>	3
Task Name	<div><div>2</div><div></div><div>1</div><div>2</div></div>	5	Task Name	<div><div></div><div>1</div><div>1</div><div>1</div></div>	3	Task Name	<div><div></div><div>1</div><div>1</div></div>	2
Task Name	<div><div>2</div><div></div><div>3</div><div>2</div><div>3</div><div>3</div><div>1</div></div>	14	Task Name	<div><div></div><div>1</div><div>2</div><div></div><div>2</div><div>2</div><div>1</div><div>2</div><div>1</div></div>	11	Task Name	<div><div></div><div>1</div><div>2</div></div>	3
Task Name	<div><div>2</div><div></div><div>2</div><div>2</div><div>3</div><div>3</div><div>3</div><div>3</div></div>	18	Task Name	<div><div></div><div>1</div><div></div><div>1</div><div>3</div><div>2</div><div>2</div><div>2</div></div>	11	Task Name	<div><div></div><div>1</div><div>2</div><div></div><div>1</div><div>2</div><div>2</div></div>	8
Task Name	<div><div>2</div><div></div><div>1</div><div></div><div>1</div></div>	4	Task Name	<div><div></div><div>1</div><div></div><div>1</div><div>1</div><div>1</div></div>	3	Task Name	<div><div></div><div>1</div></div>	1
Task Name	<div><div>2</div><div></div><div>3</div><div>2</div><div>3</div><div>3</div><div>2</div></div>	15	Task Name	<div><div></div><div>1</div><div>2</div><div></div><div>1</div><div>2</div><div>2</div><div>1</div><div>2</div></div>	3	Task Name	<div><div></div><div>1</div></div>	1

Competitive reviews of Ratings also foster improvement

Windows vX.X.XXXX.X
iOS vX..X | Android vX.X

Our Product

92

54

Competitor

Windows vX.X.X.XX | iOS vX.X.X
Android vX.X.X-XX

Example Task 1*	NA	Example Task 1	2
Example Task 2	13	Example Task 2	14
Example Task 3 that goes on two lines	19	Example Task 3 that goes on two lines	12
Example Task 4**	16	Example Task 4***	10
Example task 5 possibly with way tinier font	32	Example task 5 possibly with way tinier font	31
Task 6 with two lines	8	Task 6 with two lines	4
Example Task 7 on two lines	1	Example Task 7 on two lines	1
Example Task 8 on two lines	3	Example Task 8 on two lines	12

*Task 1 was absorbed into Task 2, so is not rated separately
**This could be another note that you want to make

***This could be another note that you want to make

Frequently Asked Questions

Q: Why do we specify the Target User?

A: In order to reduce the variance around how raters would interpret the PURE rating for a given step. This can vary greatly by user.

Q: Why do we only look at the Fundamental Tasks?

A: In order to have a consistent baseline measure **and** to force the team to prioritize what matters most.

- You can choose to apply PURE to other tasks, even all tasks, if you plan to re-score and compare later.

Q: Why don't we deviate from the Happy Path?

A: In order to have a consistent baseline measure and to show that we're rating "our best shot" at solving a user problem.

- Once this is done, you can choose to apply a PURE rating/score to any flow, but it probably shouldn't be part of an official PURE score for the product

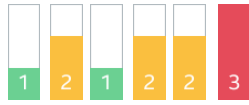
Q: Why do all raters watch the same thing?

A: Unlike Heuristic Evaluation, where you are looking for a wide number of problems, here we are trying to get a score we can reliably count on.

- If raters don't see the same thing, they will undoubtedly vary, not based on how hard it was, but based on what they happened to see

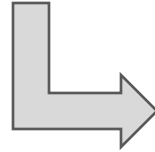
Leverage your observations!

Task 5: Install Browser Extension



11

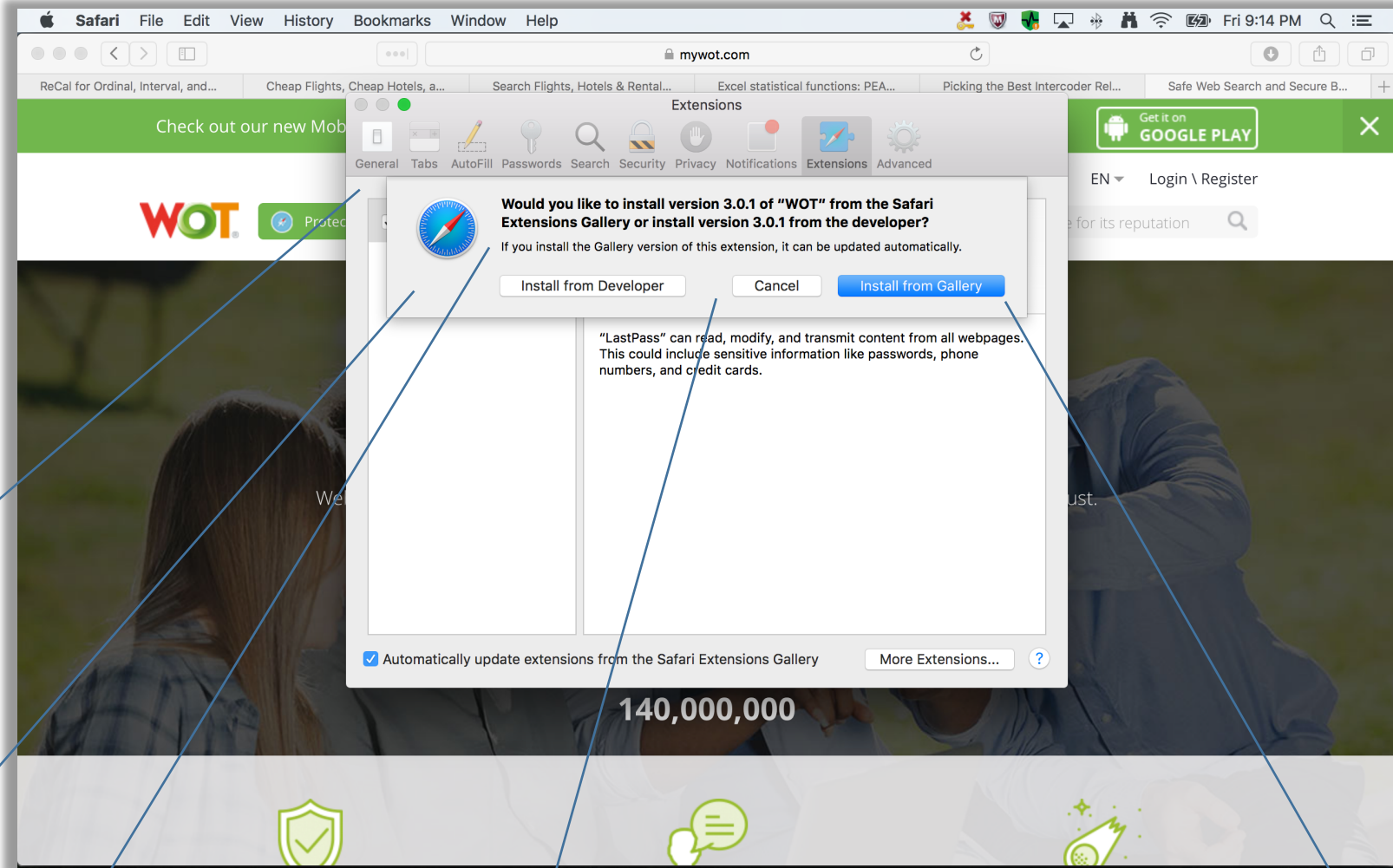
Why this got a 3:



This Extensions tab from Safari settings comes out of nowhere after the previous step (unexpected)

The dialogue box appears at the same time as the Safari settings tab, partially obscuring its content and masking the context the dialog is related to.

The language used here is difficult for the target user to fully understand without significant effort. (At least a benefit is explained, however.)



There are three choices, not uniformly spaced (so looks sloppy). Most problematic, it's likely not clear to this user type what "Cancel" does at this point without some cognitive effort.

After selecting the default button "Install from Gallery" both the dialogue and the Safari setting disappear and it appears not to have done anything (this issue is technically part of the next step)

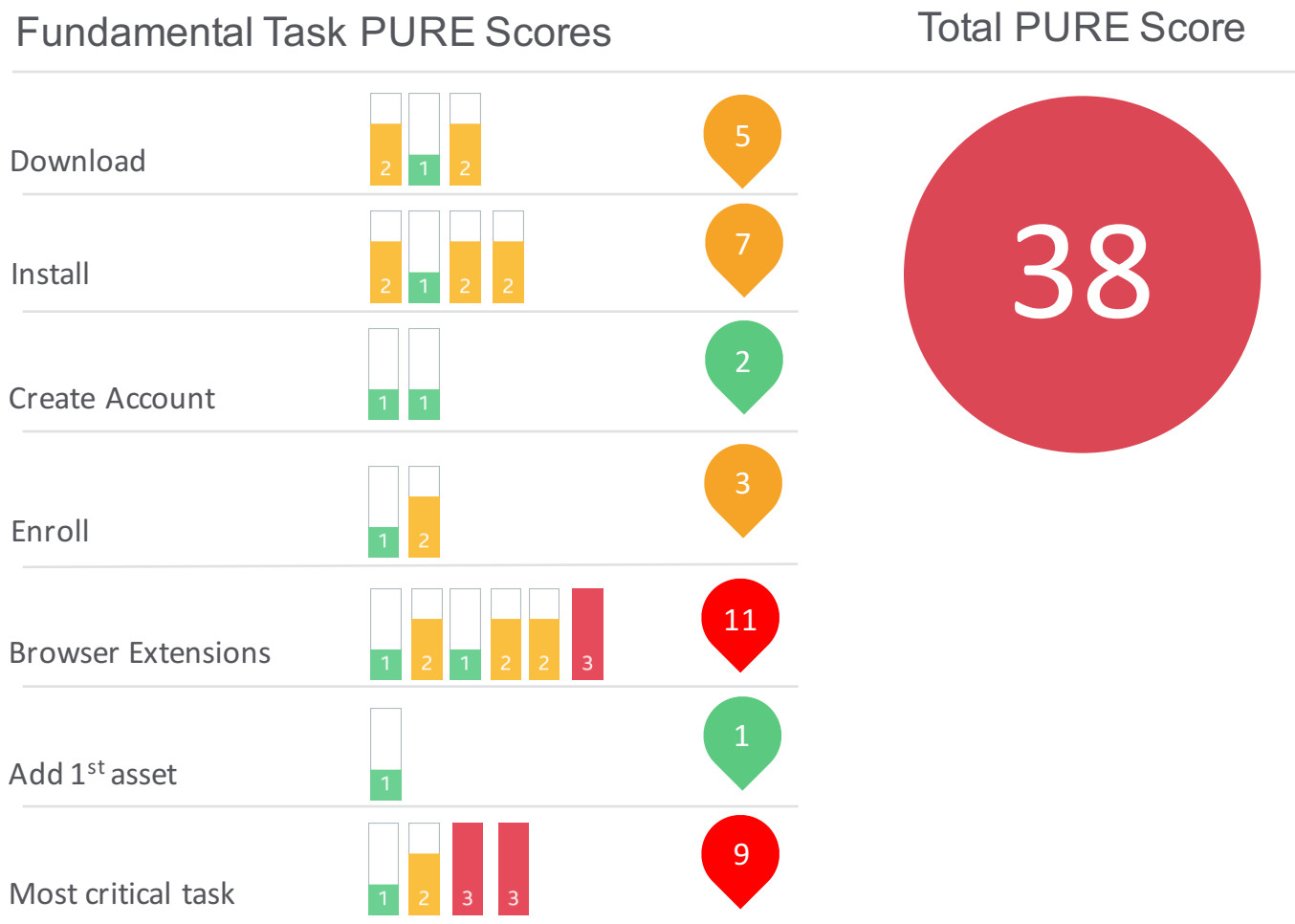
Future Improvements to Measuring UX

- PURE is only one measure of User Experience (focused on Ease)
- Other measures we are working on:
 - Measuring Content and Visual Design Guideline Compliance
 - Simple characterizations: ✓ ! ✗
 - More PURE-like ratings by task
 - Measuring “Efficacy” or “Effectiveness” or “Usefulness”
 - NPS as a proxy
 - Domain-specific measures
 - Measuring emotions in the user experience
 - Galvanic Skin response
 - Micro-facial encoding
 - Image-based emotional self-reported instruments
 - Journey analysis: the delta between the current and ideal journey

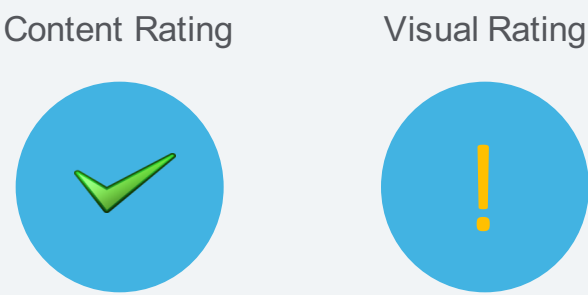


Example Product – UX Scorecard

USABILITY



APPEAL



USEFULNESS (NPS)



Summary and Closing Remarks

- PURE is not a substitute for quantitative measures of usability/UX
- Conducting PURE *and* measuring usability can extend validity further
- Qualitative studies strongly inform PURE Ratings
- Business leaders are obsessed with metrics, so **give them to them**
- End result: more improvements to UX and usability

Resources at <http://www.xdstrategy.com/PURE>

- Spreadsheet for capturing raters' scores
 - Excel
 - Google sheets
- Deck with artwork for creating PURE Score slides
 - PowerPoint
 - Keynote
- More info about PURE:
 - Original CHI2016 paper and presentation documents
 - Google Email Discussion Group
 - Leave survey feedback about PURE