IBM User Research and Design

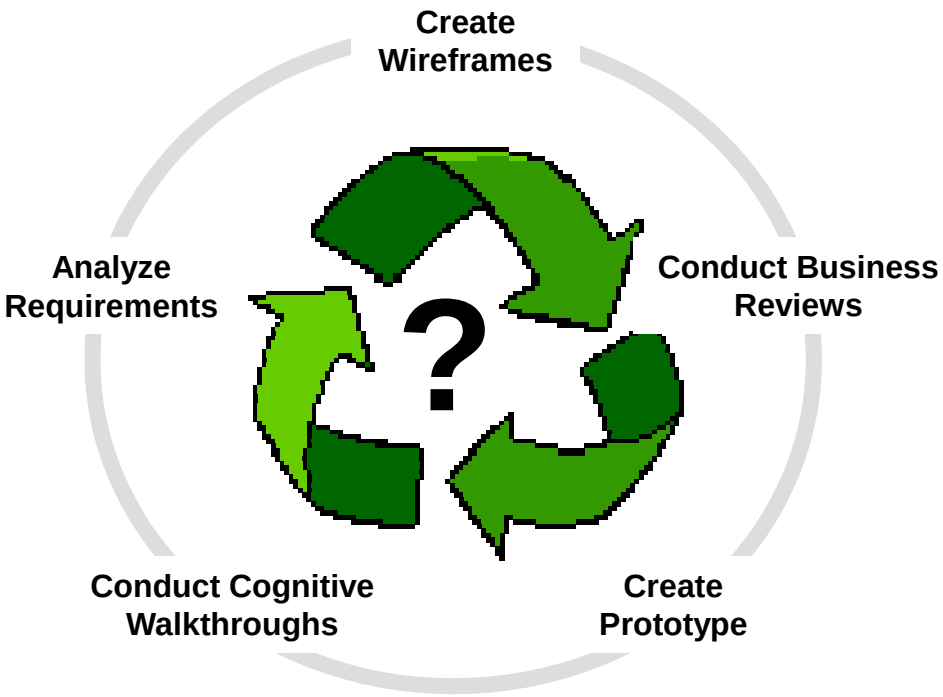
When is a User Interface Good Enough?

Usability Ratings in UI Design

IBM INTERACTIVE

Martin Fracker, Ph.D.
Michael Heck, Ph.D.
George Goeschel

Sooner or later, every UI design team faces the problem of knowing when to stop iterating: when is the user interface good enough?



Iterative user interface design is widely practiced throughout industry

- Subjective user ratings are commonly used to assess usability
- Using subjective usability measures to know when a UI is good enough often proves challenging
  - Typically collected using 5-point Likert rating scales (Likert, 1932)
  - Results are usually reported as mean ratings

But two questions are often difficult to answer:

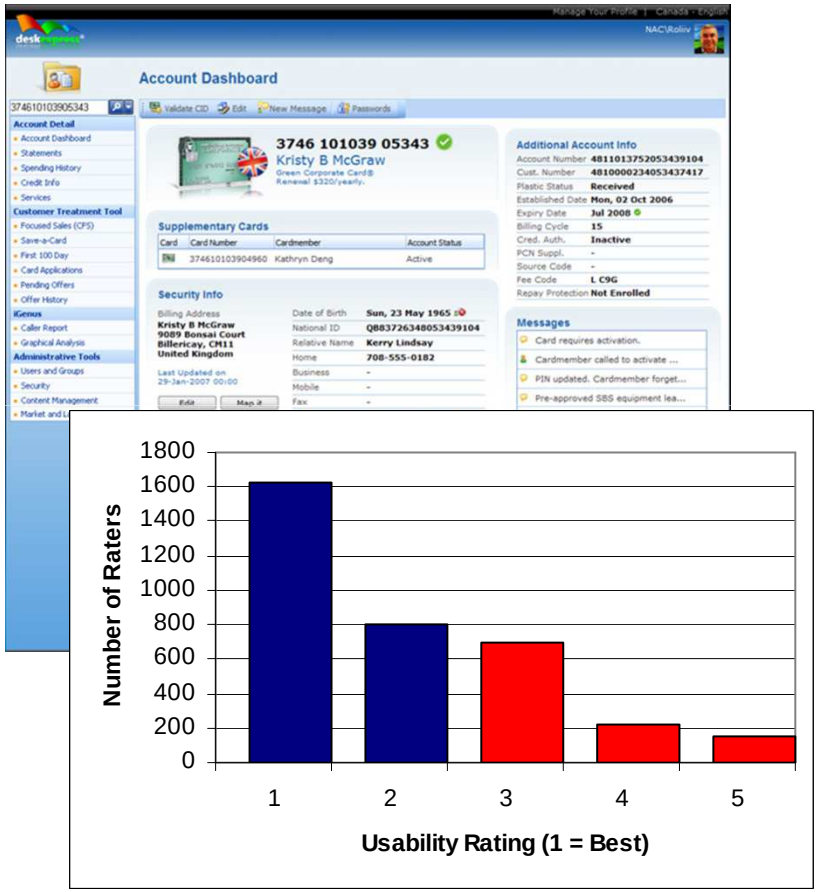
- When are these mean ratings sufficiently good?
- What are these ratings actually measuring?

A client once asked us, "How do you know the ratings are not good just because the user interface is new and different?"

To answer these questions, we focus on two issues:

- Usability Scaling
- Construct Validity

Usability Scaling answers the question: When is the usability good enough?



A Bank builds a prototype of a new Customer Servicing Workstation and User Interface

- Ten Customer Service Reps try it and rate its usability on a 5-point Likert scale (1 is the best score, 5 is the worst)
- Because the mean rating is 1.9, the Bank decides to build and deploy the new workstation

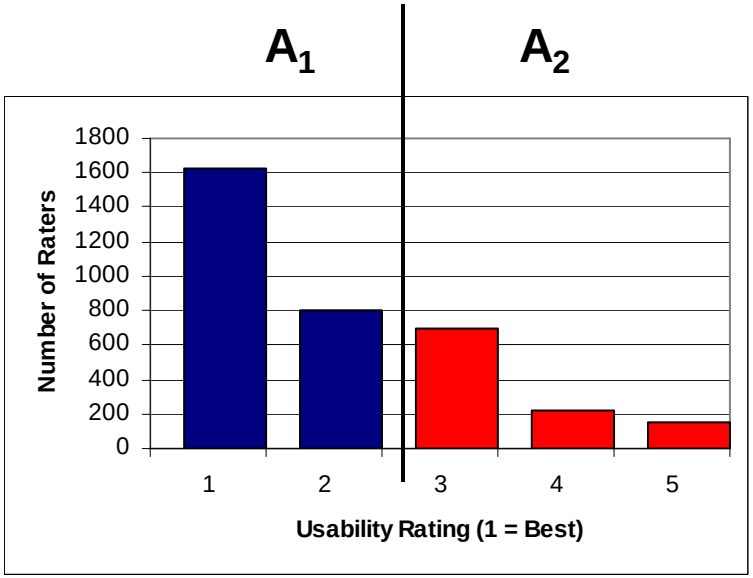
Over the next year, the bank collects satisfaction ratings from 3,500 CSRs

- The mean rating is 1.99
- But over 30% rated the workstation $\geq$ 3.0

The Bank would not have built the new workstation if they had known that over 30% of CSR's would be less than satisfied with it.

Is there a way to scale usability ratings in terms of the probability that future users will select a rating of 2 or better?

To scale usability ratings in terms of the likelihood that future users will select a rating of 2 or better, we can use the probability distribution of Likert-type ratings.



Y_1 = 1630 + 800 = 2430
Y_2 = 700 + 220 + 150 = 1070
p' = $Y_1/(Y_1+Y_2)$ = 2430/3500 = .69

U = 100(.69) = 69

Fails to meet usability requirement of U ≥ 90

Responses to a 5-point Likert rating scale follow a multinomial distribution (Clason & Dormody, 1994)

- Let the random variable X_i be the number of raters who choose each rating i, $1 < i < k$
- X_i has a multinomial distribution with parameters n, $p_1, p_2, \ldots, p_{k-1}$.
  - k is the number of points on the rating scale
  - n is the number of raters
  - p_i is the probability a rater will select rating i

We can now define a new random variable Y_j which is the number of ratings that fall into regions A_1 ($i \leq 2$), and A_2 ($i > 2$).

- Y_j has a **binomial** distribution with parameters n, p
- We estimate p from $Y_1/(Y_1+Y_2)$
- Our usability scale score is simply U = 100p

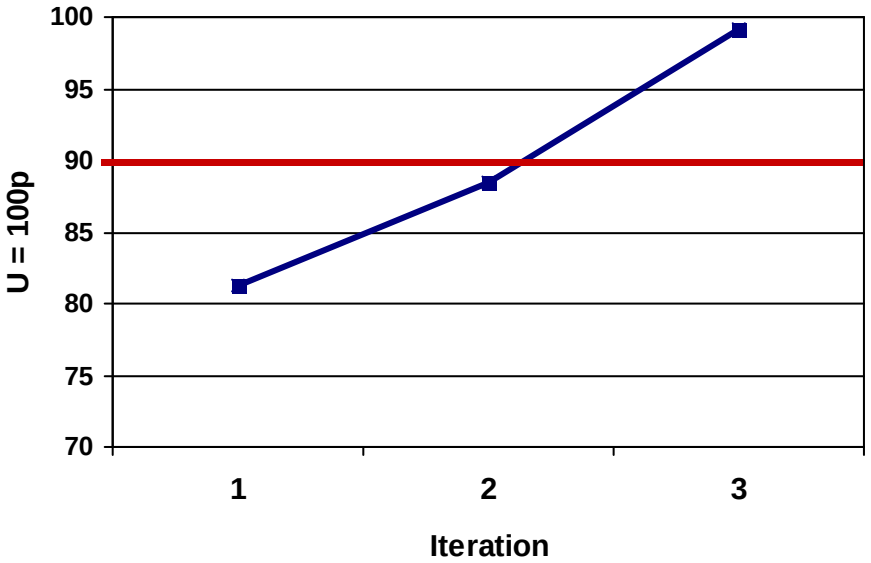
We can then establish a Usability Requirement that U = 100p should meet or exceed 90

Interpretation: *Our goal is that at least 90% of the user population is likely to rate the usability of the new workstation as a 2 or better.*

Study 1: We put the usability scale score to work on a project to design a new Customer Service workstation for an international financial services company.

The Usability Scale Score improved with each iteration, from 81 on the first iteration to 99 by the 3rd iteration.

If we had stopped iterating on iteration 2, we would have terminated the design process prematurely.



Three Design Iterations

- UI to support making by-phone payments and adjusting late payment fees
- Each iteration ended with a cognitive walkthrough
- 13 Customer Service Reps (CSRs) participated in each cognitive walkthrough (14 were scheduled)
- CSRs participated in pairs – in order to control costs
- At the end of each walkthrough, each CSR independently responded to a series of 9 or 10 Likert-type rating scales

Rating items were designed to be diagnostic rather than generic indicators of satisfaction

- Items focused on specific features of the UI design
- CSRs indicated whether they agreed with positive statements about each feature

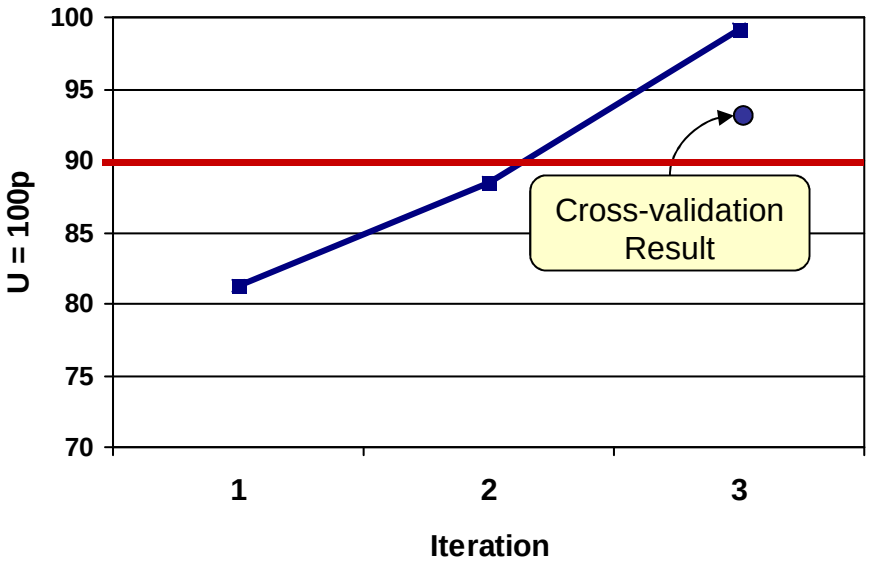
It was easy to find the transaction that caused the late fee.

I liked that you could add a bank account without leaving the payment screen.

1 = Strongly Agree **5 = Strongly Disagree**

Study 2: We applied the same basic Iteration 3 design to a new scenario in order to cross-validate that the usability really did meet or exceed the 90% criterion.

Using the same basic UI design but with a different scenario, the cross-validation study obtained a Usability Scale Score of 93



Cross-Validation Scenario

- Customer is calling in response to a promotion
  - Not sure whether to accept the promotion
  - Or upgrade to another card product
- CSR reviews advantages of each option with customer who then decides to accept the promotion

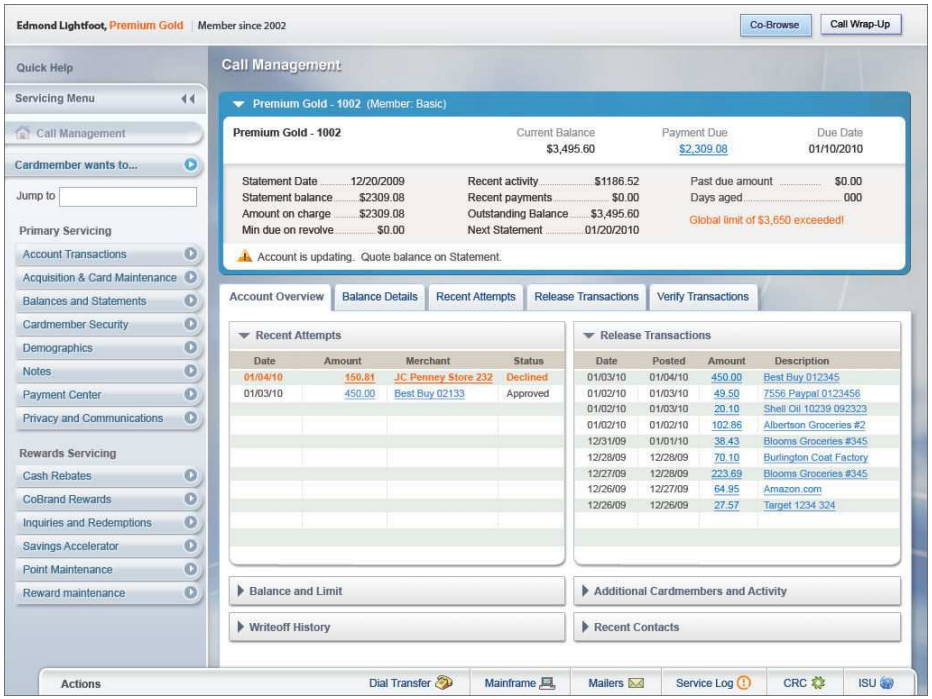
Cross-Validation Methodology

- 15 CSRs participated in one of two focus groups
- Focus group rather than cognitive walkthrough format used at client's request to reduce costs
- Facilitator presented the UI to participants and explained how they would interact with the screens

At the conclusion of the focus group, each participant completed 14 diagnostic Likert-type rating scales

Overall Results: U = 93

Study 2: We evaluated our hypothesis that usability ratings reflect at least three dimensions of user interface design: functionality, content/data, and layout.



Functionality: I like what I can do on this screen

Content/Data: This screen presents useful data elements

Layout: I like the way this screen is organized

1 = Strongly Agree **5 = Strongly Disagree**

A user interface has at least three dimensions:

- Functionality – What can users do?
- Content/Data – What information do users see?
- Layout – How is it all organized?

We hypothesized that

- Each dimension contributes to the perceived usability of a user interface
- These dimensions can vary from screen to screen
- That overall usability ratings will reflect the "goodness" of these dimensions across all screens

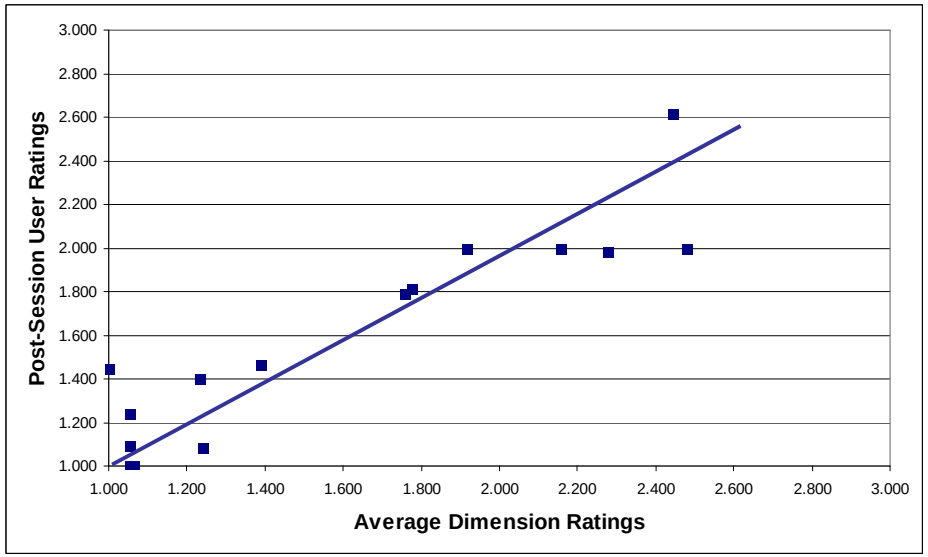
We predicted that

1. Screen by screen ratings of functionality, content, and layout would predict our post-session diagnostic ratings
2. All three dimensions would be needed to fully account for overall usability ratings

The results of Study 2 supported Prediction 1 but not Prediction 2 – the screen-by-screen functionality, content, and layout ratings were highly intercorrelated, but they strongly predicted post-session ratings.

Intercorrelations among Functionality, Content/Data, and Layout ratings

	Content	Function	Layout
Content	1.000		
Functionality	0.916	1.000	
Layout	0.924	0.990	1.000



Results

✗ *Prediction 2*: The functionality, content/data, and organization ratings were highly intercorrelated

- Any one of the three was sufficient to predict post-session ratings
- For further analysis, we took the average of the three ratings

✓ *Prediction 1*: The screen-by-screen averaged ratings predicted post-session usability ratings

- On logical grounds, we constrained the linear model to have a zero intercept
- As expected, the linear relationship was strong: $F(1,14) = 52.93$, $p < .00001$, $R^2 = .79$
- Combined dimension ratings accounted for 79 percent of the variance in the usability ratings

What These Results Mean

- Our post-session usability ratings do seem to reflect users' perceptions of the individual screens
- But participants probably failed to distinguish among functionality, content/data, and layout
- Future tests of our hypothesis should try to vary the three dimensions independently