10.1 Making Metrics Matter

By Todd Zazelenchuk

Collecting usability and other design related metrics has become a hot topic in recent years as usability has become more of a mainstream concept for many organizations. The consumer software industry, the world of home appliance design, and institutions of higher education, are just a few examples where organizational leaders have found themselves enamored with the collection of metrics as a way of helping their organizations 'move the needle' in their product design efforts. Collecting quantitative measures of a product's performance, however, is only part of the equation. In order for usability metrics to stand a chance of influencing the future direction of a product, several criteria must be met. Without them, the effort may resemble a successful academic exercise, but will most likely fail to have the desired impact on the product's direction. The following case study illustrates one such example where usability metrics were successfully collected, but their ultimate impact was limited.

10.1.1 OneStart: Indiana University's Enterprise Portal Project

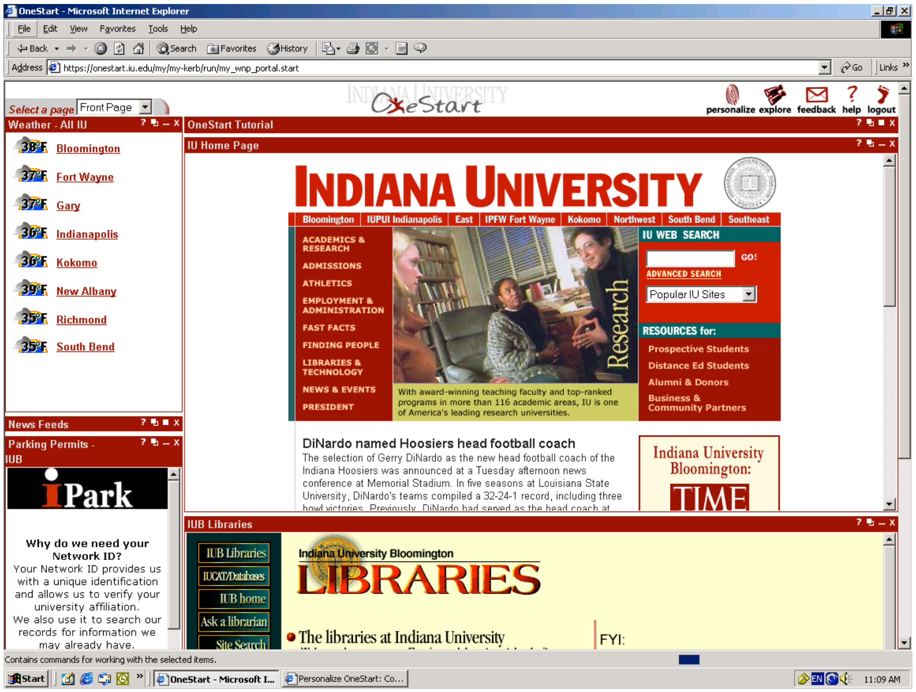
Indiana University (IU) embarked on its enterprise portal project in the year 2000 with design research and iterative prototype development leading the way. Technically, the project had begun two years earlier with the publication of an information technology strategic plan for the university (McRobbie, 1998). This plan identified a broadening base of information consumers who were becoming increasingly tech-savvy, and whose expectations for convenient, quick access to information and services were rapidly expanding. While the plan never actually mentioned the word 'portal', it effectively described the need for what would become *OneStart*, a "next generation" enterprise portal responsible for providing a full range of university services to over 500,000 students, staff, faculty, and alumni (Thomas, 2003).

Integral to the IU Strategic Plan was *Action 44*, the requirement for a user-centered design approach to all information technology projects. From 1995-2003, Usability Consulting Services, an internal consulting group based within IU's University Information Technology Services (UITS), supported project teams in the design and evaluation of their numerous software development initiatives. Known as the User Experience Group today, this team has since contributed significantly to the successful technologies delivered by UITS and Indiana University. In the case of the OneStart project alone, more than a dozen research studies have been conducted on various aspects of the portal over the past seven years, including usability testing, user surveys, and focus groups.

In 2000, not yet able to test any designs of its own, the *OneStart* team began with a comparative evaluation of some existing web-based portals. Three portals (MyExcite, MyFidelity, and My Yahoo) were evaluated with a sample of student and faculty users. The emphasis was largely on navigation and personalization tasks (selecting content for display, arranging a custom layout, changing background themes and colors). From this study, the team gained insights into many of the design elements that made portals of that era either easy or difficult for users to interact with and comprehend.

By early 2001, the team had a working prototype of *OneStart* in place, and the next phase of testing began. There were several motivations for the next round of research. At the most basic level, the team wished to understand how users would react to their university information and services being consolidated into the new portal environment. We anticipated that users may be confused about the relationship between the new portal and the traditional home page of the IU website. Further motivation involved a desire to learn whether the content organization and personalization features of the portal were both usable and useful for the target population of users. Finally, the author was selfishly motivated to complete his dissertation related to the topic of measuring satisfaction as an attribute of usability. Together, the combination of these motivating factors led to an empirical study with the following goals:

- Identify the major usability problems associated with the portal's navigation and personalization features in order to help direct the next iteration of *OneStart.*
- Establish usability benchmark data (comprising effectiveness, efficiency, and satisfaction metrics) for the core tasks currently supported by the portal in order to allow comparison with future design iterations of *OneStart.*
- Investigate the theoretical questions of whether certain methods of administering user satisfaction surveys have an impact on the ratings themselves, and whether correlations between efficiency, effectiveness, and satisfaction exist for portal users.
- Identify why users rate their satisfaction with the portal the way they do (i.e. what are the contributing factors of a portal experience to users' satisfaction or frustration with the product).



With award-winning teaching faculty and top-ranked programs in more than 116 academic areas, IU is one of America's leading research universities.

DiNardo named Hoosiers head football coach

The selection of Gerry DiNardo as the new head football coach of the Indiana Hoosiers was announced at a Tuesday afternoon news conference at Memorial Stadium. In five seasons at Louisiana State University, DiNardo's teams compiled a 32-24-1 record, including three bowl victories. Previously, DiNardo had served as the head coach at

Why do we need your Network ID?
Your Network ID provides us with a unique identification and allows us to verify your university affiliation.
We also use it to search our records for information we may already have.

The libraries at Indiana University

Contains commands for working with the selected items.

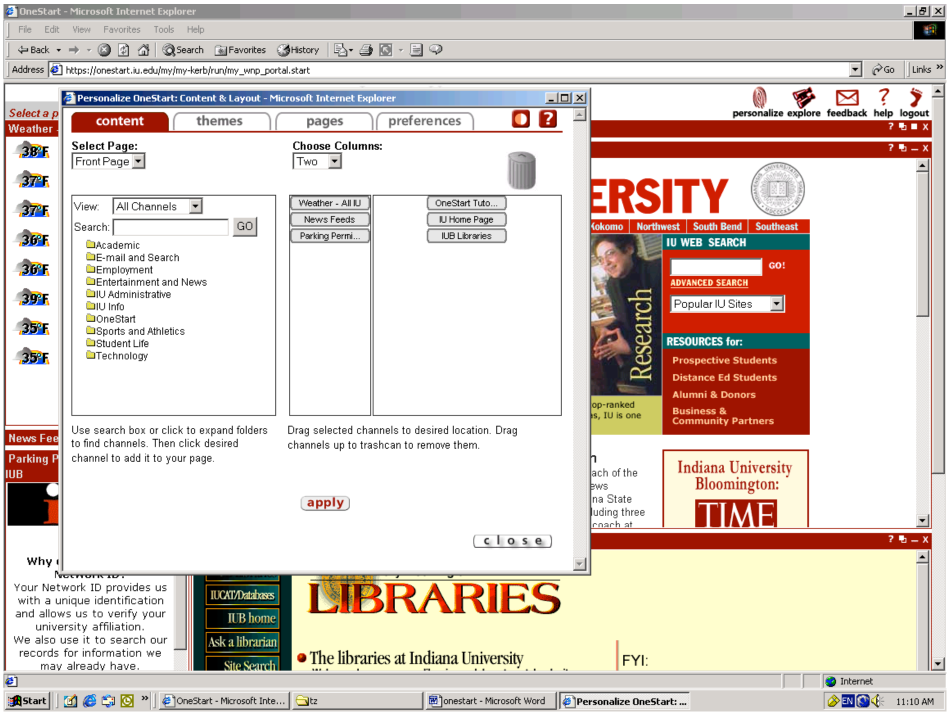
Figure 10.27. Indiana University’s OneStart portal (August, 2001)

10.1.2 Designing and conducting the study

To achieve the goals outlined for the research, a usability lab study was designed and conducted with a sample of 45 participants representing the student portion of the overall *OneStart* target population. This was a much larger sample than the lab normally recruited for formative evaluation studies, but the desire to collect certain metrics and apply inferential statistical methods made it necessary. Had it not been for the dissertation related questions, a smaller sample and descriptive statistics would have sufficed.

The study applied a between-subjects, one-variable, multiple conditions design (Gall, Borg, & Gall, 1996), in which the 45 participants were distributed across three groups of 15, each of which encountered the same portal design and core tasks to be performed, but experienced different conditions for rating their satisfaction levels with the product.

The tasks for each subject included a combination of information retrieval and personalization tasks. Information retrieval tasks consisted of locating “channels”, or groups of content to be added to the subjects’ portal page. Personalization tasks required the user to change the look and organization of their interface (e.g. screen color, layout, add content, etc.) (Figure 10.28).



Personalize OneStart: Content & Layout - Microsoft Internet Explorer

Address: https://onestart.iu.edu/my/my-kerb/run/my_wnp_portal.start

Select Page: Front Page

Choose Columns: Two

View: All Channels

Search:

Use search box or click to expand folders to find channels. Then click desired channel to add it to your page.

Drag selected channels to desired location. Drag channels up to trashcan to remove them.

Figure 10.28. The personalization window of the OneStart portal (August, 2001)

A traditional two-room, mirrored glass lab facility was utilized with the researcher moderating the study from the test room, while the participant worked through assigned tasks in the participant room. The ISO definition of usability (ISO 9241-11, 1998), comprised of the three attributes, *effectiveness*, *efficiency*, and *satisfaction,* was used as the basis for the metrics collected. For *effectiveness*, a rubric was established to judge whether task performances were scored as a pass or fail. A stopwatch was used to measure the attribute of *efficiency*, the time spent per task in minutes and seconds.

The third attribute, *satisfaction*, was collected using two different instruments, the After-Scenario Questionnaire (ASQ) and the Post-Satisfaction Survey of Usability Questionnaire (PSSUQ) (Lewis, 1995). The ASQ consisted of three questions asked after the completion of each task. The PSSUQ consisted of 19 questions asked after the completion of the entire study. Both questionnaires utilized a 7-point scale (1=strongly agree, 7=strongly disagree) that was reversed prior to data analysis.

10.1.3 Analyzing and interpreting the results

We analyzed our qualitative data looking for high frequency patterns of behavior that might suggest inherent problems with the design. We found several, along with problems that were lower frequency, yet potentially severe in their impact on the user experience. Once this analysis was complete, we prioritized the problems based on frequency and our subjective ratings of severity, to help prioritize the order of presentation in our final report.

The most frequently demonstrated problems involved personalization activities, with key problem areas including tasks such as creating a custom page for personal content, changing the color of a page, and viewing the completed page. These were all considered to be rather serious problems at the time, given the importance that the team believed personalization features would have on user adoption of the portal.

For the quantitative data collected, we calculated descriptive statistics for effectiveness, efficiency, and satisfaction to share with the project team. We evaluated effectiveness by calculating the mean values of task completion for each task, as well as the mean and standard deviation for all tasks combined (M=.731, SD=.238). Efficiency (mean time per task) was presented both for individual tasks as well as for the full set of tasks (M=467.4s, SD=145.0s). Satisfaction was evaluated by reversing the scale values and computing the mean post-test PSSUQ scores for each group and for all participants (M=5.1, SD=1.1).

The theoretical questions for the study were analyzed further using SPSS to discover moderate to high correlations existing between effectiveness, efficiency and satisfaction (see Table 10.9). The different satisfaction collection methods revealed no significant difference between methods (Zazelenchuk, 2002).

ISO Attributes of Usability	**Correlations found**
Satisfaction Effectiveness	(-.593, *p* <.01)
Satisfaction & Efficiency	(-.452, *p* <.01)
Effectiveness & Efficiency	(-.394, *p* <.01)

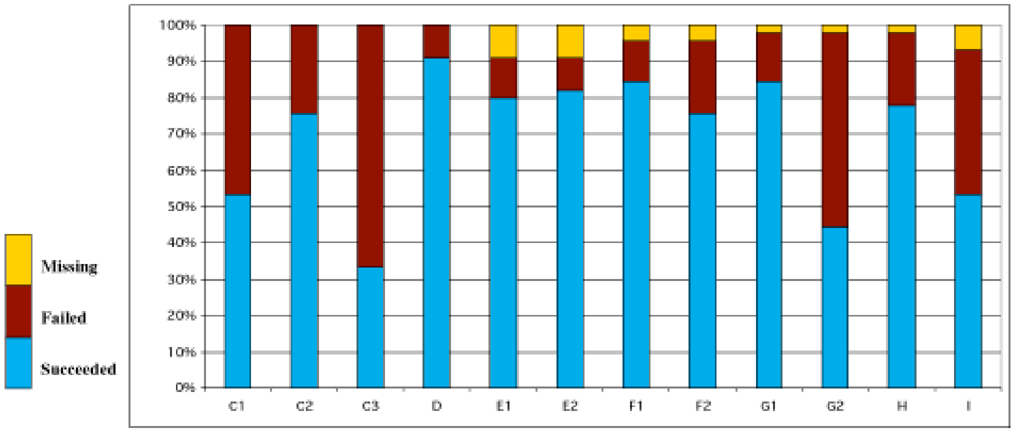
Table 10.9. Correlations between usability metrics

10.1.4 Sharing the findings and recommendations

The findings from the study were compiled and reported to the *OneStart* design team in both a written report and a presentation supplemented with video highlights of the most frequently occurring, highest severity issues. While this author has rarely compiled test session video highlights for presentation, this study represented an exception due to the large sample size. The impact of viewing tightly edited sequences of multiple participants (often 10 or more) demonstrating the same unanticipated behaviors, certainly drove the message home to the design team for many of the findings.

The quantitative data representing effectiveness and efficiency were shared with the design team on a per task basis (see Figures 10.29 and 10.30). Given that there was no significant difference discovered between the three conditions applied in the study, users' satisfaction measures were presented as an average post-task score for all 45 participants.

Task Success & Failure Rates

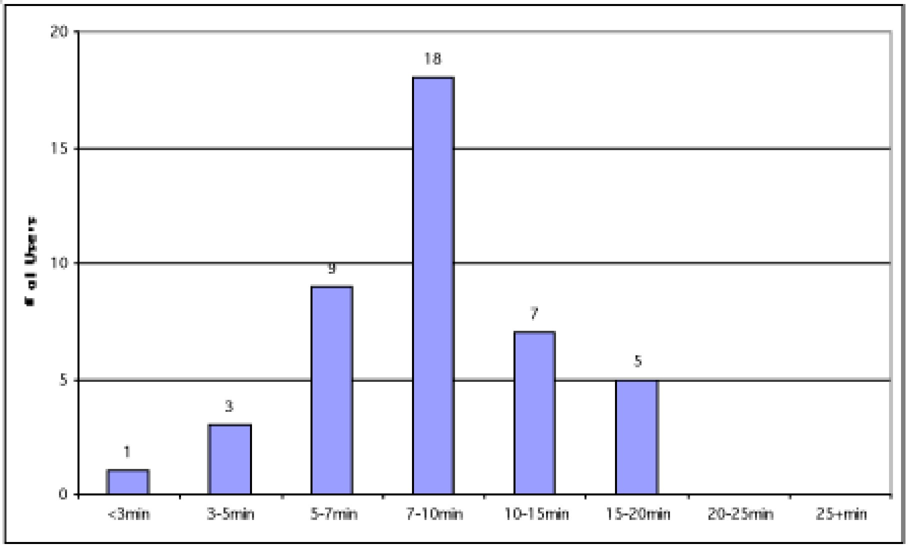


C1 - Create MyStuff page	**E1** - Add INSITE	**G1** - Create MyCC
C2 - Add 4 channels	**E2** - Print schedule	**G2** - View MyCC
C3 - Change Colors	**F1** - Create CI page	**H** - Add Bookmark/Folder
D - Add IDS channel	**F2** - Add 3 channels	**I** - Edit & Move Bookmark

Figure 10.29. Task Success & Failure Rates

Task G Times

Create a Custom Channel and View it



Min 2:09 **Mean 8:51** Max 18:55

Figure 10.30. Mean Time per Task

From a practical perspective, the most actionable data collected from the study were the qualitative findings revealed in the prioritized problem lists, and supported by the video excerpts in the summary presentation. A total of seven qualitative themes were identified representing users' rationales for their ratings of satisfaction with the portal (Zazelenchuk and Boling, 2003), and in 2005 were part of Educause's recommended reading list for the Top Ten Issues In Information Technology (Educause, 2005).

The quantitative metrics were also shared with the design team, but a reliable frame of reference for their interpretation was lacking. Had the initial competitive evaluation of existing portals been conducted with the goal of establishing benchmarks for certain tasks, those results could potentially have represented a meaningful frame of reference for the analysis. Without those baseline scores, however, the metrics collected in this study were limited to answering the academic questions associated with the author's dissertation, and providing an initial benchmark for future evaluations of the portal.

10.1.5 Reflecting on the impact

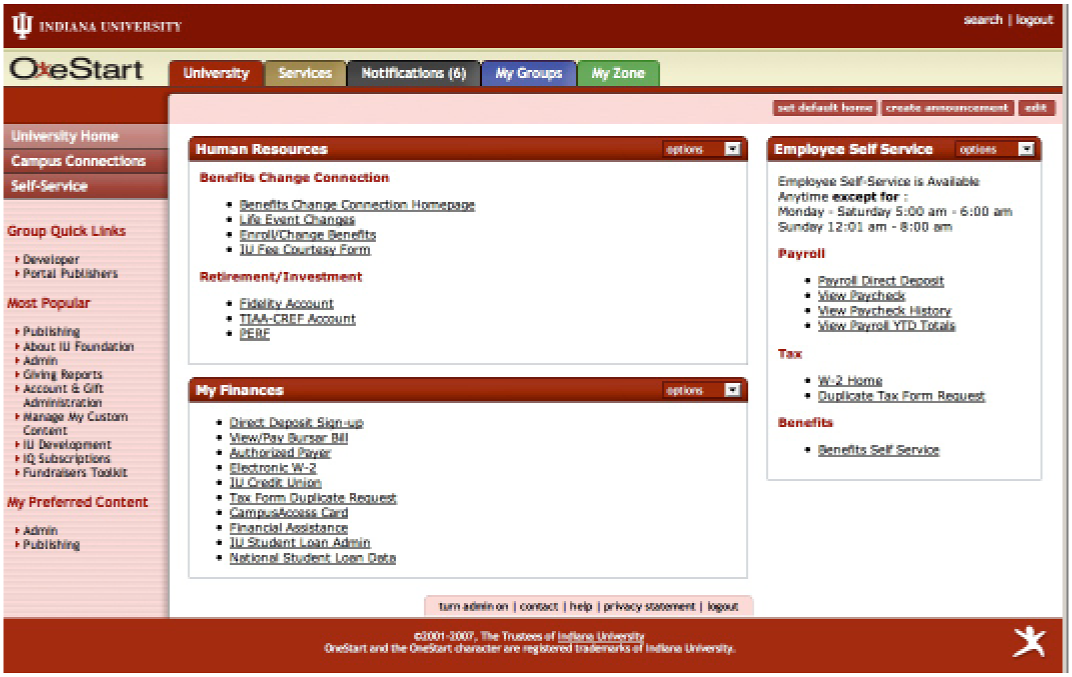
Six years after the original study, and four years after the author's last direct experience with *OneStart*, an update from the design team provided additional insights into the challenges associated with making usability metrics matter. The metrics collected in the 2001 study had provided negligible long-term value. While they successfully addressed the academic questions associated with the original study, their practical impact on the actual product was low. There were two primary reasons for this; both represent important considerations for today's organizations as they race to institute a metrics-driven usability process.

Usability metrics only provide value when there is a frame of reference. Without it, teams are left to wonder whether 80% task completion is a good score, if 85% may be necessary, or just how much of a problem it is when someone "takes 30 seconds to locate the popcorn command the first time they use a microwave oven"? When there is a clear plan in place for reliable, repeated measures to be collected in the future, an effective frame of reference can be established and valuable comparisons and learning may begin.

In the case of *OneStart*, the metrics collected in the 2001 study represented the first attempt at measuring the usability of the portal. As a result, the numbers lacked a meaningful reference point and were much less actionable than the qualitative findings from the study.

Usability metrics are most reliable and informative when the tasks being measured represent core tasks that will likely persist throughout the life of the product. Spending time collecting metrics on anything but a product's core tasks contributes to the "frame of reference" problem by constantly measuring new tasks for the first time.

In the case of the original *OneStart* study, the tasks measured were largely feature-driven. In other words, they represented the tasks that the portal supported at that time, rather than those that were truly core tasks for the product over the long term. Moreover, those feature tasks have since been found to be less important than once imagined. Web server log data (another valuable usability metric), representing the actual usage of *OneStart*'s personalization features over the past four years, have revealed that only 12% of users have ever visited the portal's personalization features. This has helped lead the team to rethink their emphasis on personalization options in the latest 2007 release (see Figure 10.31) by scaling back personalization to focus on simplicity, clarity of organization, and navigation. Given this change in direction, it suggests that collecting repeated measures of the original personalization tasks would not have been the best use of their time.



Employee Self-Service is Available
Anytime **except for** :
Monday - Saturday 5:00 am - 6:00 am
Sunday 12:01 am - 8:00 am

©2001-2007, The Trustees of Indiana University
OneStart and the OneStart character are registered trademarks of Indiana University.

Figure 10.31. Indiana University's *OneStart* portal (2007) (http://onestart.iu.edu)

10.1.6 Conclusion

The OneStart case study represents a common example of where the efforts expended to carefully measure a product exceeded the returns. It reminds us that collecting usability metrics should be kept in perspective; they are a means to an end, where the "end" is the improvement of your product or process. By ensuring that you have in place a frame of reference to help you interpret your metrics, and that you restrict your focus to core tasks that can be revisited in future evaluations, you are more likely to produce metrics that matter.

10.1.7 Acknowledgements

Thanks to Dr. Philip Hodgson, Dr. Helen Wight, James Thomas, and Nate Johnson for their critical feedback on earlier drafts of this chapter.

10.1.8 References

McRobbie, M., Architecture for the 21st Century: An information technology strategic plan for Indiana University. 1998, Indiana University: Bloomington, IN.

Thomas, J., Indiana University's Enterprise Portal as a Service Delivery Framework, in Designing Portals: Opportunities and Challenges, A. Jafari and M. Sheehan, Editors. 2003, Information Science Publishing: Hershey, PA. p. 97-120.

Gall, M.D., W.R. Borg, and J.P. Gall, Educational research: An introduction. 6th ed. 1996, White Plains, NY: Longman.

ISO, ISO 9241-11: Ergonomic requirements for office work with visual display terminals (VDTs); Part 11 - Guidance on usability. 1998, International Standards Organization. p. 22.

Lewis, J.R., IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. International Journal of Human-Computer Interaction, 1995. 7(1): p. 57-78.

Zazelenchuk, T.W., Measuring satisfaction in usability tests: A comparison of questionnaire administration methods and an investigation into users' rationales for satisfaction. Dissertation Abstracts International, 2002. 63(05A): p. (UMI No. 3054425).

Zazelenchuk, T.W. and E. Boling, Considering user satisfaction in designing web-based portals. Educause Quarterly, 2003. 26(1): p. 35-40.

Educause, Recommended Readings of the Top-Ten IT Issues. www.educause.edu/ir/library/pdf/ERM0566.pdf, 2005.

10.1.9 Biography

Todd Zazelenchuk is a user experience researcher at Intuit in Mountain View, CA. He earned his Ph.D. in Instructional Technology from Indiana University in 2002. Prior to the consumer software industry, Todd worked in academia (Indiana University) and consumer goods (Whirlpool Corporation), gaining insights to both the value and challenges of applying usability metrics to the product design process.