Output vs. Outcome: Measuring Usability in a Business to Business Context

G.M. Cooiman
University of Twente
PO Box 217, 7500 AE Enschede
the Netherlands

g.m.cooiman@student.utwente.nl

ABSTRACT

Advancements in web technology require a more extensive approach for measuring usability in business to business environments. A hybrid questionnaire, consisting of quantitative and qualitative questions is conducted among experts in the field to compare two kinds of metrics. The results show that outcomes-based metrics are a more valuable kind of metric to measure usability compared to output metrics. This research was conducted in a small set of sectors, and could lead to other results in different sectors. To completely prove that outcomes-based metrics measure usability better, research has to be conducted in different sectors as well.

KEYWORDS

Output metrics, outcomes-based metrics, usability

1. INTRODUCTION

Developments in web technology enabled applications to become accessible through the internet which accommodated a major growth of available web applications. The evolved web technology made it possible for applications to become more interactive. To retain an optimal usability for the user of the application, user behavior needs to be closely monitored in order to keep a satisfactory user experience.[13]

Advancements in web technology also had an effect on the software development process. Continuous improvement is necessary to keep up with developments in methods and techniques. Stakeholders are actively involved during the software development process[14]. Requirements from stakeholders are often based on qualitative information. This information is originating mostly from claims from customer representatives. These claims cannot easily be quantified due to the lack of available or interpretable information. A way to create the necessary information is by using web analytics tools that collect this data[13].

Typical web usage analytics are focused on output metrics. Output metrics address what was produced or provided. For example: direct user feedback, amount of orders and conversion rate[6][11]. In case of usability, metrics could be the amount of finished tasks or a satisfaction score from the user. These metrics are widely used within the b2c markets[11]. However, research in the business to business (b2b) market doubt the adequacy of output metrics for measuring usability[15]. In contrast to positive b2c experiences which are assumed to be engaging, robust, compelling and memorable, positive b2b experiences are trouble-free and reassuring and based on reducing customer effort[15]. Output metrics cover the direct outputs of the system, such as the conversion, the amount of clicks, retention[6][11]. These metrics are not covering the aspects that are necessary for a positive b2b experience, as direct outputs do not cover measurements in the amount of effort a user has to finish specific tasks[14]. Earlier research was publicized, but mostly on b2c. A possible reason for this is that b2b is researched but not publicized.[8]

Output metrics are considered to be too high level, and not covering what actually matters in business to business (b2b) usability[8]. Research has been conducted to propose more adequate approaches for measuring usability. This research suggested a conceptual approach by using outcomes-based metrics. Outcomes-based metrics describe the benefit of the service to the customer and what was changed or accomplished as a result of the service.[15]

By means of the outcomes-based metrics, specific usage scenarios can easily be quantified, such as how long it takes to complete a certain task. These quantifications could be used to substantiate the qualitative claims from the stakeholders on usability[13]. Providing this actionable information improves product decisions and design changes[13][14].

New research has to be conducted to empirically prove that outcomes-based metrics are better suited for measuring usability.[15]

2. RELATED WORK

This research aims to verify that outcomes-based metrics are more valuable in measuring usability than output metrics. The literature review starts with finding out how usability is currently described. Scopus gives more than sixty thousand results on the term “usability”. Extending this query with “construct” narrows it down to around twelve hundred results. This research aims at measuring usability, so “measurement” is added to the query to get more accurate results. This brings the results back to one hundred and twenty three items. These results are filtered on relevance to the subject.

The most common and simplified definition of usability is “*ease of use*”. Ease of use was described by Eason as “*the degree to which users are able to use the system with the skills, knowledge, stereotypes and experience they can bring to bear*”. This definition emphasizes skills, experience, and knowledge of the user.

The follow up ISO definition of usability is: "*The extent to which a product can be used by specified users to achieve*

specified goals with effectiveness, efficiency, and satisfaction in a specified context of use."[1] The emphasis is here on effectiveness, efficiency and satisfaction. Effectiveness describes the percentage of use compared to its potential. Efficiency covers the required effort for the specified purpose. Satisfaction represents a subjective user satisfaction. Dimensions of satisfaction consist of the liking and feelings about the product, and measurements relating to attitude and perception toward the product. Compared to Eason, this definition puts more focus on the system instead of the user. The system should provide a certain representation to provide the user with the proper effectiveness, efficiency and satisfaction.

Later research[10] defines usability based on five characteristics, namely learnability (how easy and fast a user learns a new system), efficiency (the user efficiency once the system has been learned), memorability (how easy infrequent users can use the system), reliability (the frequency and impact of errors) and satisfaction (liking of the product, attitude measurement, feelings about the product). Nielsen agrees on efficiency and satisfaction, but puts more focus on the processes involving users, such as learnability and memorability.

Not all researchers agree on satisfaction being part of the usability construct. A cluster analysis on usability characteristics, compiled of the ISO definition and Nielsen definition of usability, has been performed to uncover how users think about the integrality of characteristics to the usability construct. The characteristics associated with "satisfaction" were separated from other usability related characteristics, and therefore classified as not closely related to usability[9].

More recent literature is stipulating that the usability construct is "dead". Usability is considered to be a umbrella construct, which can be put as broad and diverse. Usability in this case is considered to be vague and loose, with characteristics that are challenging to communicate and measure relations between. A presented way to move forward could be by unbundling the usability construct and replacing it with well defined constructs[12]. However, extensive commentaries reject this conclusion to look for alternative constructs. The international debate asks for a constructive way forward. They recognize the importance of the issues but do not agree on the usability construct being dead. Their proposition is to look for weaknesses and identify strategies to move forward and mitigate the issues while harmonizing the usability construct[3].

The literature shows that there is no widely supported consensus on the contents of the usability construct. But not only the construct and characteristics of usability are critiqued, but also the measures related to the characteristics. Recent research critiques the adequacy of the measures to cover the complexities of usability in a b2b environment by using output measures. This research proposes a more strategic measurement approach by using outcomes-based metrics to not only capture the results, but also emphasize on the context of the results. This provides deeper understanding into usability. However, more extensive research is required to prove these conceptual claims[15].

3. RESEARCH QUESTION

Previous research concluded that outcomes-based metrics need to be empirically researched to find out if this approach allows better usability measurements.

The following research question will be answered:

To what extent are outcomes-based metrics better suited for measuring usability in a b2b environment compared to output metrics?

4. METHODOLOGY

This research answers the research question by conducting a questionnaire. Proven measures of the characteristics of usability are used to compare output metrics with outcomes-based metrics. The comparison of these measurements need to be scored on the dimensions of good measurement[5]. The semi-structured questionnaire will be filled in by specialists at a company called Centric[4]. Centric is a company that is involved in multiple sectors which enables this research to cover those sectors.

The goal of this empirical research is to find out which type of metric is considered better suited for measuring usability. This is an inductive approach: data is collected to base a theory on. A combination of quantitative and qualitative answers are collected to be able to compare the metrics and form a theoretical basis on the given scores. The quantitative results are analyzed and result in a score on validity, reliability and practicality. The qualitative feedback will result in a list of pros and cons of outcomes-based metrics, posing a qualitative foundation of the score and a basis for discussion.

4.1 Characteristics

According to research, the usability construct is constructed on five characteristics, namely: learnability, efficiency, memorability, reliability and satisfaction[10]. As discussed in Section 2, empirical analysis among users showed that satisfaction is considered not closely related to usability, and will therefore not be included in this research[8]. The remaining four characteristics are used to define appropriate measures.

4.2 Measures

The four characteristics of usability are learnability, efficiency, memorability and reliability. These characteristics are used to base measures on.

4.2.1 Learnability

The main idea of learnability is that during the first contact with the system, the user should easily become familiar and competent. For example, when someone needs to make a declaration for the first time, the user should be able to do the steps fairly quickly. A variety of established measures for learnability exist[7]. An output metric and an outcomes-based metric will be selected.

Learnability can be measured with performance, but also with user feedback. User feedback is also widely used as a measure for usability in b2c[2][7], therefore the collection of user scores is selected as output metric, as it is a direct output of the system. Performance is measured with units of time. An established metric for performance is task completion time[10]. A goal for this metric could be that on average every task is finished within thirty seconds. The outcome metric is the task completion time.

4.2.2 Efficiency

The main idea of efficiency is that when users have interacted with the system and established some experience, the user can perform the task faster or perceives it as better. For example, when a user makes multiple declarations, does this have a positive effect on the time to complete the task. This can be

perceived as learnability over time[10]. Also for the output metric efficiency can be measured, as the user can perceive the task as better after practice over time.

4.2.3 Memorability

The main idea of memorability is when users do not interact for a longer period of time, and than have to do the same task, can they re-establish their proficiency. For example, if you have to create a new declaration after three months not using the system, what effect does this have on the task completion time[10]. This measure is comparable to learnability and efficiency, only there is the influence of elapsed time between subsequent tasks. The output metric is the score over the time between two subsequent tasks. The outcomes-based metric is the task completion time over the time between two subsequent tasks.

4.2.4 Reliability

The main idea of reliability is measuring how many users make errors, what the severity is of these errors, and how easy the users recover from the errors. For example, if you are trying to create a declaration but the save button is not working, do you try it again or do you reload the page.[10]

The errors are logged in the server. These logs are the output of the system, and therefore the amount of errors is used as output measure. Output measures cannot measure the effects of the error, as it is not a direct output. The goal of the outcomes-based metric is to measure the amount of users that complete the task when confronted with an error, over the total amount.

4.3 Measurements

To be able to compare the output metrics with the outcomes-based metrics, dimensions need to be defined in order to objectively determine which one of the metrics is better. Good measurements[5] are judged on validity, reliability and practicality. These dimensions are described below.

The validity of metric is based on whether it measures what it is supposed to measure. For example, when the characteristic revolves around task success, collecting completion times is not covering what needs to be measured.

The reliability of a metric is related to consistency. Does the metric give a consistent result? For example, a weight scale is consistent if it constantly overestimates your body weight by a fixed amount of weight.

The practicality of a metric can be measured as how easy it is to understand and administer it. For example, data that is already being collected does not require any further data collection. As well as understandability, the results of the metric should provide a clear view on what is important.

4.4 Analysis

The questionnaire consists of twelve quantitative and twelve qualitative questions. The quantitative questions are based on a five point Likert scale which is used to score the comparison between the output metric and the outcomes-based metric. The quantitative results of the questionnaire are analyzed to conclude a statistical significant difference between the two types of metrics. The qualitative questions result in an explanation of the given score. The explanations together will form a distilled list of reasons why output or outcomes-based metrics are better suited for measuring usability, and form a basis for discussion.

5. RESULTS

The questionnaire resulted in eleven responses. From these eleven respondents, eight are product manager, two are project manager, one is business analyst and one is business development manager. The respondents operate in three different sectors; five in human resources, three in retail and three in the public sector solutions.

5.1 Quantitative results

A standard statistical report of the results from the quantitative questions is shown in Table 1.

	N	Mean	St. Dev.	St. Error Mean
validity	44	3,95	1,200	,181
reliability	44	3,34	1,293	,195
practicality	44	3,36	1,382	,208

Table 1: Statistical report on questionnaire results

The questionnaire consisted of four metrics which are all scored on validity, reliability and practicality. These four metrics filled in by eleven different people resulted in 44 scores (N=44) for validity, reliability and practicality.

The mean values are all three bigger than 3. A value of 3 means that the metrics are scored as equal. An analysis will prove with a level of statistical significance if the mean value is higher than 3. Before such a test can be performed, analysis has to conclude what the distribution of the data is. A normality test is performed to conclude whether or not to use parametric tests. This test concluded that the results are not normally distributed. Therefore a nonparametric test is performed to test whether the results are statistically significant. There are two methods, the sign test on the median and the Wilcoxon rank sum test. As the latter one requires two independent samples with equal variance which is not applicable in this case, the sign test on the median is used. Table 2 represents the results of this test.

Null Hypothesis	**Sig.**	**Decision**
Validity = 3	0,000	Reject the null hypothesis.
Reliability = 3	0,026	Reject the null hypothesis.
Practicality = 3	0,230	Retain the null hypothesis.
The significance level is ,05.		

Table 2: Results of nonparametric test on the significance of the results.

The test hypothesizes that the median is equal to 3, at a level of significance of 0,05.

This hypothesis is rejected for the validity and reliability, but retained for practicality. This means that there is no statistically significant difference between output and outcomes-based metrics for practicality. For validity and reliability there is a statistically significant difference, meaning that outcomes-based metrics are considered better than output metrics.

The amount of responses could have influenced the normality test. Therefore, also a test on the significance is performed while assuming normality. A parametric test on the difference between the control value 3 (which implies equality) and the mean values of the validity, reliability and practicality will be executed. A One-Sample T-Test is used. This test is presented in Table 3.

	Test Value = 3					
			Sig.		95% CI	
	t	df	(2-t)	MD	Low	Up
validity	5,277	43	,000	,955	,59	1,32
reliability	1,749	43	,087	,341	-,05	,73
practicality	1,745	43	,088	,364	-,06	,78

t: test statistic, df: degrees of freedom, Sig.(2-tailed): two tailed significance, MD: Mean difference, 95%CI: 95% Confidence interval for the difference.

Table 3: Results of parametric test on the significance of the results.

The parametric test calculates a 95% confidence interval for the mean value being 3. The Low and Up column define the interval for which the Mean Difference is considered equal to the Test Value. This test concludes that there is a statistically significant difference in the score between output and outcomes-based metrics on validity. Reliability and practicality are considered equal for both types of metrics.

5.2 Qualitative results

The answers on the qualitative questions explained the given score. These answers are distilled into two lists. The list of reasons why outcomes-based metrics are better than output metrics is presented in Table 4. The list of reasons why output metrics are better than outcomes-based metrics are presented in Table 5.

Reasons
This actually shows the effect of learning the system and directly adds value in customer talks.
Practicality depends on the available assets. If measuring software is in place to collect the necessary extra data, outcomes-based metrics would definitely be better.

Table 4: List of reasons why outcomes-based metrics are better than output metrics.

Table 4 indicates that outcomes-based metrics are giving a more complete view on usability. The practicality is considered to be dependent on available assets. In case the necessary assets are available, outcomes-based metrics are considered better.

Reasons
Validity of outcomes-based metrics can be biased due to factors influencing the results that have little to do with usability.
Outcomes-based metrics are less consistent over time due to potentially changing circumstances on the client's side.

Table 5: List of reasons why output metrics are better than outcomes-based metrics.

Table 5 indicates that the validity and reliability of the outcomes-based metrics are debatable. Some respondents doubt the accuracy of measuring how long a task is taken. A selection of given statements is given below.

"Time only gives limited input about how easy users experience a task because there is no further context about how long the task should take, etc.". The adequacy of time measuring how users perceive the task is debatable. Short tasks can be perceived harder than long tasks.

"Time is very unreliable. It will be influenced by factors that have no real association with effiency. Such as alt-tabbing. Or internet issues on the client. this will polute the numbers." The time a task takes is also not very accurate due to users doing other things on the side, or connection issues on the user's side

"Time is tricky to measure and not reliable. People could be doing other stuff and become distracted, have tabbed pages. What is 'time'. So while time tells us very well if learnability is achieved in a lab scenario. The measurement of time is in reality very problematic." The issue of time is mentioned multiple times. Where closed lab experiments can be free of bias by giving users only the task at hand, no distractions can influence the results of lab experiments. This is undoable in real life, and therefore does not allow accurate measurements.

6. CONCLUSION

Over the years technology has become more complex. Web technology is changing all the time, and is growingly interactive with the user. To be able to find improvement points, metrics are used to measure usability. However, research on usability is merely focused on the b2c market. The b2b market requires metrics that enable more extensive usability analysis therefore more effective ways to measure usability in a b2b environment need to be researched. Literature review implied the potential of outcomes-based metrics. This research investigates the potential of outcomes-based metrics compared to output metrics. Information is gathered in a questionnaire conducted among experts in the field. This questionnaire resulted in quantitative data on the different metrics and additional qualitative information. Parametric analysis of the data concluded that when a normal distribution is assumed, a statistical difference can be identified for the validity scores. The normality test showed that no normal distribution was applicable. The nonparametric analysis showed that when no normal distribution was assumed, practicality was considered statistically equal, but the validity and reliability were considered better in case of the outcomes-based metrics. Both analysis were conducted due to the small amount of available data which could have influenced the distribution of the data. The results of the qualitative part showed that the practicality of outcomes-based metrics is highly dependent on environmental factors and assets. Measuring extra data and analyzing it costs a lot of time and requires tooling which can be expensive for some companies to implement, and therefore less practical. The research question is: *"To what extent are outcomes-based metrics better suited for measuring usability in a b2b environment compared to output metrics?"*. The potential of outcomes-based metrics is proven, as it does improve the validity and the reliability of the measures for usability. The practicality is considered equal, and is dependent on environmental factors.

7. DISCUSSION

The results show that the extent to which the measures measure what they are intended to measure (validity) is perceived better with outcomes-based metrics. The metrics cover more context and zoom in on the how and why of the metric, compared to output metrics that zoom in on the what. However, the use of time in the metrics is considered biased by the respondents. Research should be conducted to test whether this claim is true or not. The consistency of the results (reliability) is perceived better for outcomes based metrics. The covered context could change, or other factors

could influence the results which would not be measured in the metric and could lead to unexplained discrepancies. This is however not considered to be of more significance than for output metrics. The expectation was that the reliability would be considered less or equal compared to output metrics. An explanation of the equal score could be that those factors could as well be reflected in the output metrics, which would lead to consistency in the discrepancy. The practicality of the metrics is perceived as equal between the metric. The context that is measured in the outcomes-based metrics requires a lot more data than output metrics. This is however not perceived as significant, but highly dependent on available assets. If the data is already being collected, the effort to use the metrics is significantly lower than when all data needs to be collected. The answers on the practicality question were very divergent. Persons with different backgrounds perceive the effort to collect the data differently. It is therefore dependent on the person you ask the question too. When people do not understand the complexity of the required data collection, scores will be biased.

The respondents of the questionnaire included experts from the sectors retail, HR and public services. This is a small selection of all available markets. To be able to claim that outcomes-based metrics are perceived better, more markets need to be approached in similar research.

The research was limited by the amount of respondents of the questionnaire. The amount of respondents in this research was eleven. This amount is small, and this could have had an effect on the outcomes of this research. Further research should test the already tested markets as well as unexplored markets to be able to adequately conclude the significance of outcomes-based metrics compared to output metrics.

The results of this research could potentially be used in other fields as well. One example is agile project management. In agile project management product development is done in sprints in a specified time frame. In this sprint is a certain amount of work which represents a certain business value that needs to be delivered at the end of the sprint. This is expressed in story points. All the work on the sprint backlog needs to be finished. When the sprint is not done at the end of the specified timeframe, the team needs to justify the unfinished sprint and set up a plan to successfully finish the next sprint. This emphasis in this approach is related to output: the number of story points that is finished that sprint. But what is the goal of a development team? Burn as much points as possible, or make sure that the clients perceive the best experience? Outcomes-based metrics could mess up this approach by putting the emphasis on customer experience and usability instead of amount of story points. This could mean that the amount of story points in a sprint is cut in half, but the customer experience perception improves with 100%. Modern project management could potentially benefit from this approach, and should be researched in the future.

8. REFERENCES

[1] Bevan, N., Carter, J., Earthy, j., Geis, t., Harker, S., (2016). New ISO Standards for Usability, Usability Reports and Usability Measures. HCI 2016, Part I, LNCS 9731, pp. 268–278, 2016.

[2] Bevan, N., Macleod, M. (1994). Usability measurement in context. Behaviour and Information Technology. 13:132-145.

[3] Borsci, S., Federici, S., Milizia, A., De Filippis, M.L.. (2019). Shaking the usability tree: why usability is not a dead end, and a constructive way forward. Behaviour and Information Technology Volume 38, Issue 5, 4 May 2019, pp. 519-532

[4] Centric: https://www.centric.eu

[5] Cooper, D.R., Schindler, P.S. (2008) Business Research Methods, 10th edition.

[6] Google Analytics: http://www.google.com/analytics

[7] Grossman, T., Fitzmaurice, G., Ramtin, A. (2009). A Survey of Software Learnability: Metrics, Methodologies and Guidelines. Autodesk Research 210 King St. East, Toronto, Ontario, Canada, M5A 1J7

[8] Jamshidi, B., (2008). Web Usability in B2B Websites. User's Perspective. Lulea University of Technology. Independent thesis Advanced level.

[9] McGee, M., Dumas, J., (2004). Understanding the Usability Construct: User-Perceived Usability. Human Factors and Ergonomics Society Annual Meeting Proceedings 48(5).

[10] Nielsen, J., (2012). Usability 101: Introduction to Usability. Human Computer Interaction 2012.

[11] Rodden, K., Hutchinson, H., Fu, X., (2010). Measuring the user experience on a large scale: user-centered metrics for web application. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2395-2398.

[12] Tractinsky, N.. (2017). The Usability Construct: A Dead End?. Human-Computer Interaction 33(2).

[13] Weischedel, B. & Huizingh, E. (2006). Website Optimization with Web Metrics: A Case Study. Proc of ICEC 06, ACM Press, pp. 463-470.

[14] Yakhneeva, I.V., Agafonova, A.N., Fedorenko, R.V., Shvetsova, E.V., Filatova, D.V., (2018). On collaborations between software producer and customer: A kind of two-player strategic game. Conference on Digital Transformation of the Economy: Challenges, Trends and New Opportunities. pp. 570-580.

[15] Zolkiewski, J., Story, V.M., Burton, J., Chan, P., Gomes, A., Hunter-Jones, P., O'Malley, L., Peters, L., Raddats, C., and Robinson, W. (2017). Strategic B2B Customer Experience Management: The Importance of Outcomes-Based Measures. Journal of Services Marketing, V.31.