JUX Journal of User Experience

Vol. 20, Issue 1, November 2024 pp. 9-26

Response Instability in User Experience Questionnaires

Martin Schrepp
UX Expert
SAP SE
Dietmar-Hopp-Allee 16
69190 Walldorf
Germany
martin.schrepp@sap.com

Jörg Thomaschewski
Professor
Hochschule Emden/Leer
Constantiaplatz 4
26723 Emden
Germany
jörg.thomaschewski@
hs-emden-leer.de

Abstract

Surveys are a widely used method for evaluating a product's user experience. UX surveys include items that ask participants to express their perceptions of various aspects of UX quality including efficiency, learnability, user interface aesthetics, or joy of use. One common item format is a statement to which participants can indicate their level of agreement or disagreement on a scale. Another frequently used format is semantic differentials: Participants indicate which of two terms along a semantic dimension is closer to their perception of the corresponding UX quality. However, participants' responses to these items are non-deterministic. Non-determinism can occur for several reasons. Participants may be uncertain about the best answer or might randomly choose among reasonable alternatives. Or situational factors might lead them to unintentionally select the wrong answer option without realizing it. Even if a person's opinion about the UX quality of a product remains unchanged, their answers could vary between two evaluations using the exact same items. In our study, we investigate the magnitude of these effects and discuss their impact on the interpretation of UX survey results. Our study reveals significant response instability in UX ratings. Fortunately, this doesn't affect the overall item or scale scores, as deviations are random and have a symmetrical, neutralizing impact on scale means with a large enough sample. However, this instability does increase the standard deviation, affecting its interpretation.

Keywords

user experience, UX questionnaire, response instability, surveys, random response errors

Introduction

Questionnaires that measure the user experience of a product contain items that describe quality aspects (such as learnability, efficiency, or trust) related to the interaction between a user and a product. The items can be grouped into scales, although some UX questionnaires may only provide an overall score.

In the typical item format, participants can express their opinion by choosing one of several answer alternatives in a rating scale, for example, in the following item from the UMUX (Finstad, 2010):

This system's capabilities meet my requirements

Strongly disagree O O O O O O O Strongly agree

Another typical format is semantic differentials. These items consist of two terms with opposite meanings. Participants rate on a scale which of the two opposing terms better describes the product. An example is the following item from the UEQ (Laugwitz et al., 2008):

Inefficient O O O O O O O Efficient

However, participants may not always be sure which alternative they should choose, or some factors that are independent from their actual UX perception may influence their choice. If we ask the same participants twice, with a delay that is long enough to ensure that they do not simply replicate their previous responses, they occasionally provide different answers to the same item. This can happen even if the participants have not changed their opinion about the product. According to Zaller and Feldmann (1992), we refer to this as response instability.

Response instability is related to the reliability of scales. In classical test theory, a test score is considered reliable if a repeated measurement under the same conditions produces the same or quite similar results (Lienert, 1989). A direct approach to evaluate reliability is test-retest reliability (Horst, 1966). In this approach the measurement is repeated multiple times (usually two times) with the same group of participants. The correlation between the results of participants in different repetitions is used as a measure of reliability. Thus, a low level of response instability causes a high level of test-retest reliability. However, there is a big difference between psychometric tests that measure psychological attributes on an individual level versus UX questionnaires that always measure on the level of groups of users. UX practitioners are usually interested in the average evaluation of a target group rather than the opinion of a single user. Thus, the concept of reliability is not optimal for UX scales (Schrepp, 2020).

Test-retest reliability and response instability are related concepts, but they are used in different contexts. Test-retest reliability focuses on the stability of test scores to assess the reliability of a measurement instrument, whereas response instability focuses on the variability of individual responses to survey questions to ascertain how individuals' responses may change in different contexts or over time.

We present this study to estimate how likely response instability occurs in typical UX items. In addition, we discuss the impact of response instability on the interpretation of item and scale means, standard deviations, and confidence intervals.

What Causes Response Instability?

Why do survey participants respond differently to the same question when asked twice? An obvious reason is that they may have changed their opinions during the interim period. But there are also other factors tied to the cognitive processes involved in creating a response.

Responding to a rating scale question is a decision-making task. Because survey participants are required to make a choice among the available answer options, it is possible to apply results from psychological decision research to better understand their response behavior.

Survey participants who are asked to judge some object or stimulus on its properties need to develop a mental representation of the problem (Schwarz, 1999) by recalling knowledge from memory. In our case, survey participants were asked to judge a product on a UX quality aspect like efficiency by recalling past interaction experiences. As an example, for participants judging the efficiency of a product, relevant experiences might be long versus short wait-times for

system responses, lost information entered twice versus smooth information entry, inefficient versus efficient navigation options, and so on. However, participants will not retrieve all the relevant information stored in their memory. Instead, retrieval stops once they have gathered enough information to decide with sufficient certainty (Simon, 1956; Johnson & Payne, 1985; Gigerenzer & Goldstein, 1996). This effect is known as satisficing in psychological research. The role of this effect is also described in the psychological theory of the survey response process that distinguishes the four phases of comprehension, recall, judgement, and response mapping (Tourangeau et al., 2000) in the cognitive process of subjects answering a survey question.

Thus, the outcome of a decision process is not solely based on all available and relevant information but rather on a subset of that information. The recall of specific information depends on its importance to the participant and how recently it was acquired. Highly important information or events that created a massively negative or positive impression may be always recalled (such as losing a lot of valuable work due to a system bug). The probability of recalling other information depends on the context (for example, when a previous question activated some memories) or if the participant's attention was directed toward that information (Strack et al., 1988), for example, as part of the instructions to the survey.

In addition, other factors can influence survey responses. For example, a participant's attention may be momentarily distracted due to an interruption during answering the survey. Thus, the participant may accidentally click on a neighboring alternative instead of the intended answer without recognizing this mistake. Another example of such errors occurs due to a lack of attention in questionnaires that mix the polarity of responses, such as positive alternatives placed on different sides of the survey for different items. In such cases, participants may unintentionally select the wrong alternative if their positive impressions align with the opposite side. While these situations are not frequent, they do occur (Sauro & Lewis, 2011; Schrepp et al., 2023).

Different users naturally have varying histories of product usage, which can influence their judgments on efficiency. Additionally, users may employ the product for different purposes, leading to diverse impressions regarding UX aspects (Schrepp, 2021). Personal experience with a product also plays a role because more experienced users tend to provide higher ratings (McLellan et al., 2012). Thus, participants' ratings for a specific item depend on the experiences they stored in memory; the sum of these experiences collectively represents a participant's opinion. These more or less random fluctuations, known as response instability, are caused by the incomplete retrieval of information from memory and distractions, or fluctuations of attention and concentration, while answering the survey.

Response instability in customer satisfaction questions is investigated in several papers on marketing research (Westbrook, 1980; Torkzadeh & Doll, 1991; Lam & Woo, 1997; Dawes et al., 2020). Results showed that respondents are to some extent inconsistent concerning their satisfaction ratings if they are asked to rate the same vendor several times. The degree of inconsistency varied over these studies.

Similar results are found in studies concerning perceived brand attributes (Riley et al., 1997; Rungie et al., 2005). If participants of a study agreed that a brand has a certain attribute but were asked again after some delay, approximately only half of them agreed the second time. In such cases response instability is surprisingly high.

Structure of the Study

Our study estimates the impact of response instability on typical UX items.

Overall Concept

Our objective is to assess the stability of UX ratings. The basic idea of the empirical study was to determine variations between two ratings of the same product by the same individual. Thus, the participants rated the same products twice. Rating was done in two surveys that contained the same items from four established UX questionnaires.

There are two crucial factors we considered in our experimental setup. First, we aimed to prevent participants from simply recalling their ratings from the first survey and replicating responses in the second survey. This would result in an underestimation of response instability. Therefore, we introduced a time gap between the two surveys. Second, participants' opinions

regarding a product's UX may change over time, which can affect their ratings of UX items. This would result in an overestimation of response instability. Hence, we selected products that were familiar to participants, reducing the likelihood of changes in opinion. Additionally, we set the time interval between the two surveys to a value long enough to make it unlikely for participants to remember their previous ratings. It was still short enough to minimize the chance that they fundamentally changed their UX impression of the product.

Selection of Products

It is important for the study that participants were familiar with the products they would evaluate. The greater their experience and usage frequency, the less likely they were to alter their opinion between the two surveys about the products' UX. Therefore, we provided a list of 6 popular products. Instagram™, Netflix®, and Spotify® are products used primarily for entertainment and leisure. Moodle™, Amazon™ product search, and MS Teams™ are products used primarily to achieve specific goals. Participants were instructed to evaluate only products with which they had sufficient experience. Additionally, if possible, they were asked to evaluate one entertainment and one goal-oriented product.

Participants

We recruited participants from a course held in the winter semester of 2023 at the University of Applied Sciences Emden/Leer (Germany). Participants received some course credits for their participation.

Seventy-one students evaluated at least one product in the first survey. Four students did not provide data for the second survey. Therefore, the data from 67 students were used for the data analysis. Concerning gender, 30 reported to be male, 36 to be female, and 1 did not provide an answer to this question. The average age was 28.28 years (standard deviation of 6.54 years).

UX Items Used in the Surveys

The surveys contained four established, standardized UX questionnaires featuring different item formats and response scales of varying lengths.

The 8 items of the UEQ-S (Schrepp et al., 2017) were presented first. These items consist of semantic differentials with a 7-point answer scale, as shown in the example in the Introduction section. The UEQ-S allows calculating two sub-scales measuring pragmatic and hedonic quality and an overall score.

Following that, the 4 items of the UMUX (Finstad, 2010) were presented. These items are brief statements to which participants can indicate their level of agreement using the extreme answer options "Fully disagree" and "Fully agree," as illustrated in the example in the Introduction section. The UMUX items have mixed polarity, meaning that agreement is associated with a positive user experience for two items, and disagreement is associated with a positive user experience for the other two items. The UMUX originally had a 7-point answer scale (Finstad, 2010). But further developments of this instrument used a 5-point scale (Sauro, 2017). Because we wanted some variation in the number of response options, we adapted this for our experiment. Several studies varied the number of categories on the answer scale to investigate the impact on the results of UX questionnaires. Results showed that the length of the answer scale has only a small impact (Lewis & Erdinc, 2017), thus we did not expect this change to the original answer scale to negatively impact our results.

Next, 4 items from the VISAWI-S (Moshagen & Thielsch, 2013) were shown. The VISAWI-S is a questionnaire designed to assess the visual appeal of a user interface. The format of the VISAWI-S items is similar to that of the UMUX items, but the answer scale consists of 7 points, with the extreme endpoints "Strongly disagree" and "Strongly agree." Agreement is associated with a positive user experience for all items.

Following this, the NPS® (Reichheld, 2003) item was presented, which uses an 11-point answer scale. This instrument consists of a single question "How likely is it that you would recommend <product/service> to a friend or colleague?" and an 11-point answer scale with the endpoints "Not at all likely" and "Extremely likely." We report the NPS results on the 0–10 scale and not as scores based on the percentage of promotors (participants rating 9 or 10) or detractors (participants rating below 7) due to the limited size of our data set.

Thus, our study contains a number of UX items from several established UX questionnaires. These items are different types and contain answer scales of varying lengths. Table 1 presents the English translation of the items included in our surveys (the original German version can be found in Table A in the Appendix).

Table 1. Items Used in the Surveys

Item	Item Text
UEQ-S item 1	obstructive/supportive
UEQ-S item 2	complicated/easy
UEQ-S item 3	inefficient/efficient
UEQ-S item 4	confusing/clear
UEQ-S item 5	boring/exciting
UEQ-S item 6	not interesting/interesting
UEQ-S item 7	conventional/inventive
UEQ-S item 8	usual/leading edge
UMUX item 1	This system is easy to use.
UMUX item 2	This system's capabilities meet my requirements.
UMUX item 3	Using this system is a frustrating experience.
UMUX item 4	I have to spend too much time correcting things with this system.
VISAWI-S item 1	Everything goes together on this site.
VISAWI-S item 2	The layout is pleasantly varied.
VISAWI-S item 3	The color composition is attractive.
VISAWI-S item 4	The layout appears professionally designed.
NPS	How likely is it that you will recommend the product to a friend?

Procedure

The participating students could access the survey through the respective Moodle course. The first survey was open from November 10 to November 19, 2023. The second survey was open from November 27 to December 5, 2023. This schedule ensured a gap of at least 8 days between the first and second survey to reduce the likelihood that the participants would remember their first evaluation.

Before completing the first survey, participants entered a character string and recorded it. This string was entered into both online surveys, allowing us to identify and link data from the same participants while maintaining anonymity.

Participants chose a product to evaluate by selecting it from a drop-down menu. Participants were instructed to rate one task-related and one fun-related product, and they were asked to rate the same two products in the second survey as well. To do this, they opened the survey link twice and selected different products each time. For each product, the frequency of use ("How frequently do you use the product?") and the usage experience ("How long have you been using this product?") were requested. The first survey also contained questions about the age and gender of the participants.

The introduction of the first survey mentioned that there would be a second survey, but it was not mentioned that the same products should be rated. This prevented participants from making screenshots, noting their ratings, or paying special attention to remembering their ratings.

Our goal was to investigate response instability. Thus, we implicitly assumed that the participants did not change their general opinion about a product between the two surveys. To ensure this was the case, the second survey contained the question, "Has there been an event since the first survey that fundamentally changed your opinion of the product?" If this question was answered with "Yes" instead of "No," the corresponding data point was removed from the

analysis. This was the case for 8 of the 140 cases in which we received an evaluation of the same product by the same participant in both surveys.

The items for the evaluation of the product (Table 1) followed directly after these initial questions.

Participants were instructed to finish their first evaluation within 10 days. There was a 1-week delay after the first survey, followed by the activation of the second survey on Moodle. Participants were given another 10 days to complete their evaluation. On average, the delay between the start of the first survey and the start of the second survey was 15 days (standard deviation of 3.64 days).

Results

Time Spent on Answering the Surveys

The response time of a participant was measured as the time between the start of the survey and the click on the Submit button. Some extremely long and unrealistic times were observed. This was most likely caused when a participant was interrupted, leaving the browser window open and starting work again on the survey after a large delay following the interruption. Mean and standard deviations of the response times are difficult to interpret, so we report the median.

For Survey 1, the median completion time was 267 s (minimum 51 s, maximum 3930 s). For Survey 2, the median was 117 s (minimum 37 s, maximum 12,949 s).

Usage Frequency and Experience

We collected the following number of evaluations for the 6 products for which participants could evaluate more than one product: Instagram 23, Netflix 25, Spotify 20, Amazon 24, Moodle 26, and MS Teams 14. In 132 cases, participants evaluated the same product in both surveys.

In 96 of the 132 records, a usage frequency of at least several times a week was reported. Rare usage ("several times a year") was reported in only 5 records. In 121 of the 132 records, the reported usage experience was higher than a year, and experience of less than 6 months was reported only in 4 cases. Details can be found in Tables B and C in the Appendix.

If we assume that the probability that users changed their opinion about a product in the relatively short time between the two surveys decreased with a long usage experience and a high frequency of usage, then such changes should be unlikely for our sample. This also corresponds to only 8 participants mentioning in Survey 2 that they experienced a principal change of opinion about the UX of the product. In these cases, their data was removed because of this answer, reducing the usable surveys to 132 out of the initial 140.

To check the impact of experience with a product on the response instability, we divided our observations into two groups. The first group comprised cases in which participants reported using the product for more than 3 years (100 observations), whereas the second group included cases with less than 3 years of reported experience (30 observations). We then calculated the number of items for which the ratings differed between the two surveys for each observation (with 17 items, this number varied between 0 and 17 per participant). On average, the score was 9.7 for the group with more than 3 years of experience and 10.7 for the group with less than 3 years of experience. Notably, there was a statistically significant difference between the two groups (*t*-test, one-tail, df =128, t = -1.70, p = 0.0458). No such dependency could be found for frequency of use.

Differences in Scale Means Between the Two Surveys

The following chart compares the scores of the questionnaires (mean over all items) for the two surveys. The bars represent the overall score (mean of all items in the corresponding questionnaire) with a 95% confidence interval of this overall score. The mean and standard deviation for all products and questionnaires is reported in Table D in the Appendix.

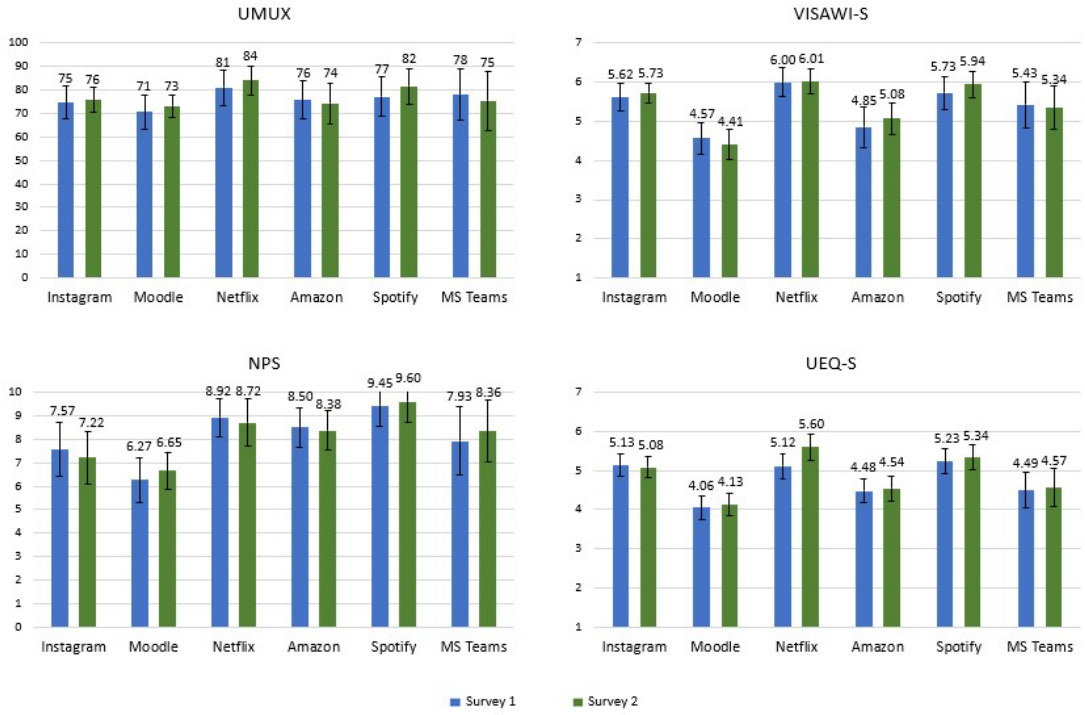

Figure 1: Comparison of the UMUX, VISAWI-S, UEQ-S, and NPS scores for both surveys (*N* = 132 participants). The y-axis corresponds to the scores according to the evaluation tools for the questionnaires: UMUX 0–100, NPS 0–10, VISAWI–S, and UEQ-S 1-7 (or -3 to +3).

None of the differences between the scores for either survey is significant (*t*-test, two-sided, paired samples, $p = 0.05$).

We also checked for significant differences between the scores of single items per product. For the UEQ-S item 5 (boring/exciting), we found a significant difference for Instagram ($t = 2.237$, $df = 22$, $p = 0.036$). For the VISAWI, none of the differences was significant. For the UMUX, we found significant differences for item 4 and Moodle ($t = 2.606$, $df = 25$, $p = 0.015$) and item 3 and Spotify ($t = 2.333$, $df = 19$, $p = 0.031$). Thus, only 3 of the 6 * 17 = 102 comparisons for single item means show significant differences, which is approximately what we can expect by pure chance on a 5% level. A Bonferoni-adjusted test would show no significant differences.

When we look at the correlations between the scale scores (aggregated over all products) between the two surveys, we get 0.76 for UMUX, 0.73 for VISAWI, 0.78 for Pragmatic Quality (first 4 items of UEQ-S), 0.74 for Hedonic Quality (items 5 to 8 of the UEQ-S), and 0.8 for the full UEQ-S (all items), which indicates an acceptable reliability of all scales.

Table 2 shows the mean and standard deviation for all items over the complete data set (the calculated mean over the 132 observations, i.e. not considering the product). The item texts can be found in Table 1.

Table 2. Mean and Standard Deviation (in Parenthesis) for All Items

Item	Survey 1	Survey 2
UMUX item 1	4.17 (0.85)	4.17 (0.79)
UMUX item 2	4.02 (0.96)	4.12 (0.79)
UMUX item 3	3.89 (1.00)	3.95 (0.99)
UMUX item 4	4.09 (0.99)	4.17 (1.01)
VISAWI-S item 1	5.56 (1.14)	5.54 (1.12)
VISAWI-S item 2	4.77 (1.53)	4.92 (1.30)
VISAWI-S item 3	5.39 (1.47)	5.45 (1.26)
VISAWI-S item 4	5.64 (1.26)	5.68 (1.19)
UEQ-S item 1	5.21 (1.28)	5.29 (1.20)
UEQ-S item 2	5.49 (1.34)	5.59 (1.18)
UEQ-S item 3	5.03 (1.41)	5.23 (1.20)
UEQ-S item 4	4.81 (1.32)	5.05 (1.38)
UEQ-S item 5	4.79 (1.43)	4.67 (1.40)
UEQ-S item 6	5.03 (1.39)	4.92 (1.32)
UEQ-S item 7	4.13 (1.67)	4.23 (1.57)
UEQ-S item 8	3.92 (1.65)	4.04 (1.42)
NPS	8.06 (2.57)	8.08 (2.48)

Only for item 4 of UEQ-S could a significant difference be detected between the mean scores of both surveys (*t*-test, two-sided, $t = -2.093$, $df = 131$, $p = 0.038$). Only 1 of the 17 tests is significant on a level of $p = 0.05$, which can be expected by chance. If we tested with Bonferoni-adjustment ($p = 0.05$/number of tests), none of the differences would be significant.

Overall, we conclude that there are no significant differences in the item or scale scores between the two surveys.

Response Instability

What is the impact of response instability on our typical UX items?

Table 3 shows each item and information concerning the deviations in the ratings between the two surveys. The columns have the following meanings:

- Equal: Number of cases in which the ratings in both surveys are identical
- Increase: Number of cases in which the rating in Survey 2 is better than the rating in Survey 1
- Decrease: Number of cases in which the rating in Survey 1 is better than the rating in Survey 2
- Size: Average absolute value of the deviation between the ratings in both surveys
- Corr.: Correlation between the scores of participants in both surveys

The number of cases with an identical rating in both surveys and the correlation provides insights into the amount of response instability per item. The size provides insights into how different the answers are between the two surveys. The scores that describe the increase or decrease of ratings can be used to determine if the effect is symmetrical (thus, it does not strongly affect the mean score), or not.

Table 3: Deviations Between the Two Surveys per Item

Item	Equal (%)	Increase (%)	Decrease (%)	Size	Corr.
UMUX item 1	80 (0.61)	25 (0.19)	27 (0.20)	0.42	0.64
UMUX item 2	80 (0.61)	22 (0.17)	30 (0.23)	0.45	0.65
UMUX item 3	72 (0.55)	33 (0.25)	27 (0.20)	0.54	0.59
UMUX item 4	75 (0.57)	32 (0.24)	25 (0.19)	0.58	0.53
VISAWI-S item 1	60 (0.45)	37 (0.28)	35 (0.27)	0.73	0.52
VISAWI-S item 2	43 (0.33)	40 (0.30)	49 (0.37)	0.98	0.59
VISAWI-S item 3	57 (0.43)	39 (0.30)	36 (0.27)	0.73	0.70
VISAWI-S item 4	69 (0.52)	30 (0.23)	33 (0.25)	0.64	0.65
UEQ-S item 1	67 (0.51)	27 (0.20)	38 (0.29)	0.59	0.74
UEQ-S item 2	61 (0.46)	33 (0.25)	38 (0.29)	0.66	0.73
UEQ-S item 3	53 (0.40)	33 (0.25)	46 (0.35)	0.83	0.58
UEQ-S item 4	49 (0.37)	30 (0.23)	53 (0.40)	0.92	0.55
UEQ-S item 5	58 (0.44)	44 (0.33)	30 (0.23)	0.73	0.72
UEQ-S item 6	61 (0.46)	43 (0.33)	28 (0.21)	0.74	0.67
UEQ-S item 7	43 (0.33)	43 (0.33)	46 (0.35)	1.00	0.65
UEQ-S item 8	45 (0.34)	43 (0.33)	44 (0.33)	0.89	0.69
NPS	54 (0.41)	37 (0.28)	41 (0.31)	1.04	0.78

Values in brackets are the relative frequencies of cases.

As we can see, the number of cases in which the scores for both surveys are identical is higher for the UMUX items than for the VISAWI-S, UEQ-S, and NPS items. If we aggregate this to the level of the questionnaires, it results in 58% consistency for UMUX, 43% for VISAWI, 40% for UEQ-S, and 41% for NPS. Because the type of response scale is associated with the semantics of an item, we cannot associate these differences with the different lengths of the response scales. For example, items that measured general usability aspects had a 5-point scale, whereas items that measured visual aesthetics had a 7-point scale.

Figure 2 shows the distribution of the deviations per questionnaire (the score in the first survey minus the score in the second survey). Note that the maximal absolute deviation is 4 for UMUX (5-point scale), 6 for VISAWI and UEQ-S (7-point scale), and 10 for NPS (11-point scale).

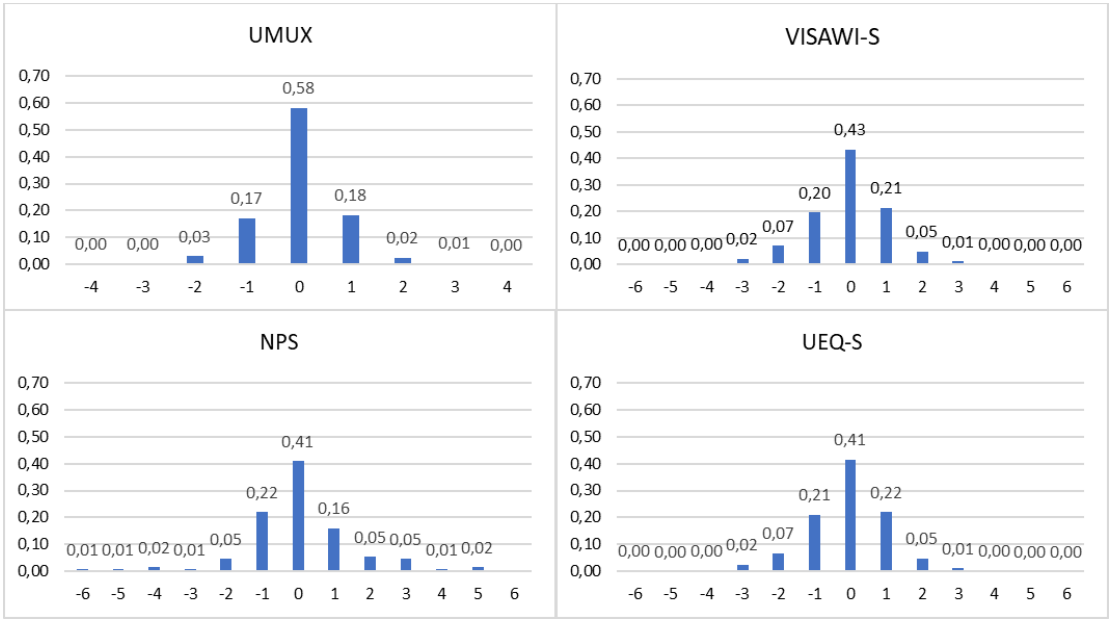

Figure 2: Distribution of the deviations over all the items per questionnaire.

Deviations are nearly symmetrical, with the case in which values in both surveys are identical (deviation 0) being the most frequent. The probability of a deviation decreases with the absolute value of the deviation.

Of course, it is a bit difficult to compare the distributions. For example, NPS is a single-item questionnaire, thus the distribution is based on 132 data points. UEQ-S contains 8 items, so this distribution is based on 8 * 132 = 1056 data points. But it is remarkable that the distribution of deviations for the UEQ-S and the VISAWI-S are nearly identical. Both questionnaires use a 7-point answer scale. The distribution of the UMUX, which uses a 5-point scale, is compared to these two distributions much more centered around 0.

We checked if the mean deviations differed between items of a questionnaire by pairwise comparisons. For the items of the VISAWI and UMUX, no significant differences were detected (*t*-test, two-tailed, $p = 0.05$). For the UEQ-S items, only 3 of the 28 pairwise comparisons showed a significant difference (items 3 and 5: $t = -2.2996$, $df = 131$, $p = 0.023$; items 3 and 6: $t = -2.100$, $df = 131$, $p = 0.038$; items 4 and 5: $t = -2.546$, $df = 131$, $p = 0.012$).

Overall, the interpretation of the deviations as more or less random fluctuations seems to be supported by our findings.

In UX studies, typically the scale scores (average over all items in a scale) are interpreted, not the item scores. The UMUX and the VISAWI-S have no subscales. The UEQ-S has subscales for pragmatic quality (items 1 to 4) and hedonic quality (items 5 to 8).

Figure 3 shows the distribution of deviations per scale. If a scale has four items, 1–4, the deviation per participant, j, is calculated as $\sum \{i = 1,\ldots,4 \mid S1(i, j) - S2(i, j)\} / 4$. The calculation is simply the deviation of the scale scores of the participants in both surveys.

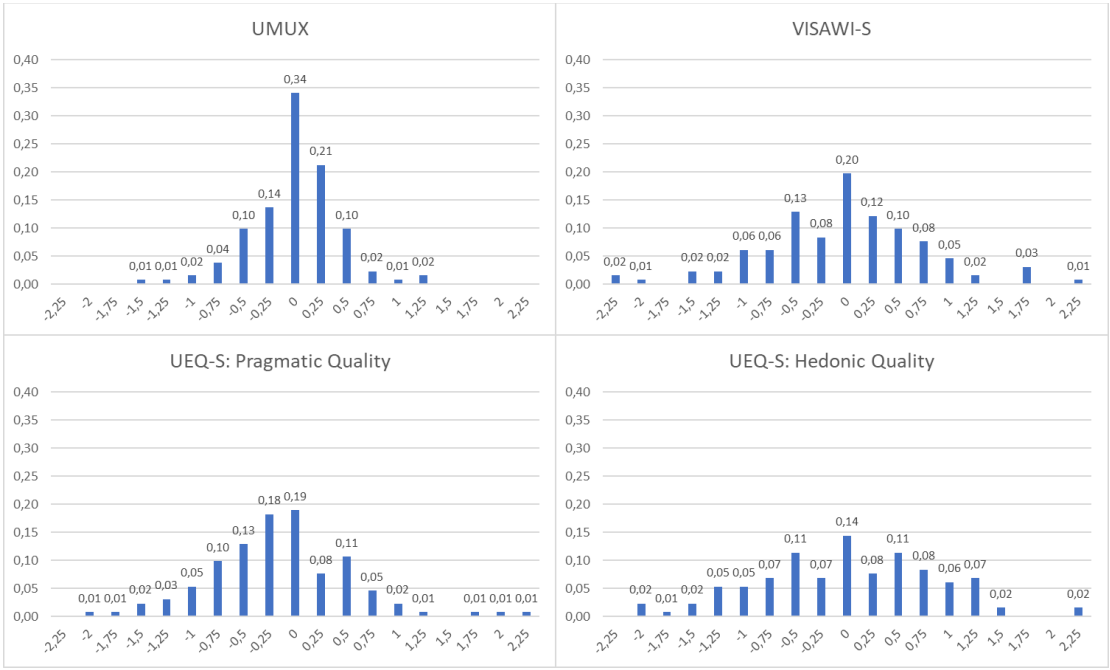

Figure 3: Distribution of the deviations per scale and questionnaire.

As we can see, the distributions are centered around 0 (though not as smoothly as the distributions for the items in Figure 2 due to the higher number of alternative values and the lower number of observations).

Another interesting question is how much instability in the ratings is contributed by single participants. If we aggregate the number of items in which the ratings between the first and second survey differed over all products rated by a participant (participants were instructed to evaluate two products in both surveys), an average of 9.74 deviations per participant results (with 17 items, this number can potentially vary between 0–17) with a standard deviation of 2.58 (min. 1, max. 15). Thus, there are high interindividual differences between participants.

A possible explanation for these interindividual differences is that the UX ratings of users are based on an incomplete retrieval process. When they are confronted with a UX item, they start to retrieve information from memory that relates to the item. But the process is stopped as soon as enough information is available to decide. What is considered to be enough, of course, depends on personal traits; for example, highly impulsive persons draw decisions faster and less accurately than other persons (Dickman & Meyer, 1988). However, there is no relation to the response times of the participants. No substantial correlation (-0.049) could be observed between the sum of response times in both surveys and the response inconsistencies of the participants.

Summary

Our results show a huge amount of response instability in UX ratings. The scores of a UX item per participant can vary if we ask them twice. However, this effect does not change the item or scale scores. The deviations are random and symmetrical. Thus, they do not impact the scale mean if the sample is big enough. This again highlights that the classical reliability concept, known from psychometric tests, is inadequate for UX questionnaires or other questionnaires that are not interpreted at the level of individual responses but always on the level of mean scores from a larger target group. Reliability in the classical sense may be low, thus response instability is high, but as long as the deviations are symmetrical, it does not impact the mean scores of a scale, which is what UX questionnaires interpret.

However, the response instability does have an impact on the standard deviations for items or scales. The standard deviation is a measure that shows how much the participants in a target group agree or disagree concerning their ratings. An increase in response instability increases the standard deviation.

And, of course, this has an impact on the number of participants needed to generate stable scale scores. The width of the confidence interval shows how accurately a scale score was measured. The smaller the interval, the more likely it is that the measured scale value represents the true value in the population of all users. For a given confidence level, this width depends on the standard deviation and the number of participants in a study (it increases with an increasing standard deviation and decreases with an increasing value of *n*). However, the standard deviation is not completely determined by response instability, thus our results cannot directly be used to predict the standard deviations or the number of participants required to generate stable results.

In Lewis & Sauro (2023), typical standard deviations of UX items are investigated, especially the dependency between scale length and standard deviation. Based on an analysis of a large sample of studies, they found that the standard deviation of UX items is approximately 25% of the scale length (scale length equals *n* - 1 for a response scale with *n* alternatives; for example, 4 for UMUX, 6 for UEQ-S and VISAWI-S, and 10 for the NPS item). Naturally, the standard deviation of scales (consisting of several items) is lower. A similar analysis (Sauro & Lewis, 2023) reported that, for several established UX questionnaires, standard deviations of scales were between 17% and 22% of the scale length. An analysis of a larger sample of studies showed an average standard deviation of a UEQ scale of 0.92 with a 95% confidence interval of [0.90,0.94] (Schrepp, 2023).

If we calculate the standard deviations for the distributions of the deviations (Figure 2), it results in 0.83 for UMUX items, 1.14 for VISAWI items, 1.16 for UEQ-S items, and 1.68 for the NPS item, which is a bit below the 25% scale length value (1 for UMUX items, 1.5 for VISAWI-S and UEQ-S items, and 2.5 for the NPS item). Under the assumption that the participants had not changed their real opinion between the first and the second survey, Figure 2 shows how much the responses of participants can vary solely based on response instability. Thus, we can interpret the values above as the amount of variability that can be explained solely by response instability. If we compare these values to the real, observed standard deviations in Table 2, we see that this factor explains a large proportion of the variance in typical item scores. A similar result can be found for the scale standard deviations. If we calculate the standard deviations for the distributions of the deviations shown in Figure 3, it results in 0.42 for UMUX, 0.82 for VISAWI-S, 0.70 for the PQ scale of UEQ-S and 0.89 for the HQ scale of UEQ-S. Thus, the deviations resulting from the response instability explain again a substantial proportion of the observed variation of scale means.

Generally, UX ratings are interpreted based on samples of users and not on an individual level. The high level of response instability shows that it is not useful to give a UX survey to the same participant several times, expecting to interpret the changes in scores as a measure of the development of a single participant's opinion. The high amount of response instability we observed in our study shows it is nearly impossible to infer any stable trend from single observations.

The concept of response instability and its possible explanations are closely related to human memory and decision-making processes. This effect will be mainly important in cases in which participants of a study judge the UX of a system retrospectively, which is a quite frequent use case of UX questionnaires in industrial practice. If a questionnaire is filled out directly after product use, for example as part of a usability test, the current usage experience will clearly dominate the ratings, and response instability will have a much smaller impact.

Conclusion

Our study investigated response instability for typical UX survey items. Participants rated the same products in two surveys with the same items. The surveys were answered with a delay of approximately 2 weeks to ensure that the participants did not simply remember and reproduce their ratings from the first survey in the second one.

Our results showed a relatively high response instability. The percentage of consistent ratings (same rating in both surveys) varied between 33% and 61% for the 17 items used in our survey. However, the inconsistencies are random fluctuations that are symmetrical, thus they do not significantly influence the mean score of an item or a scale that consists of several items. However, there is a massive impact on the standard deviation. Care must be taken when interpreting the standard deviation of items or scales as a measure of the difference in opinions concerning the UX of the product in the target group investigated. It is a measure for the variability of the observed ratings, but a large proportion of the standard deviation can be explained simply by response instability. Of course, it is unclear how the length of the delay between the two survey responses of participants influenced the result. It would be interesting to repeat the study with different delays to investigate that further.

There are limitations to our study. Our participants are students, meaning they have a higher level of education and are younger than the general population. In addition, we had to focus on products frequently used in this target group. It is, in our opinion, not very likely that the observed response instability will be much different for more representative groups of participants for other products, but of course it would be required to replicate the findings in further studies.

We used a fixed order of the items for all participants. Potentially, the order can affect the response instability, for example, due to increasing fatigue for items placed at the end. However, our survey is relatively short and such effects are not likely. In addition, a randomization on the item level would not be a good option in our case. Different items have (intentionally) a different format (statements versus semantic differentials) and different response scales and semantics; it would be quite confusing if a participant completed them in a fully random order. To check the potential impact of the presentation order on the results, we computed the correlation of the position of the item in the survey and the correlation between results of participants in the first and second survey (column "Corr." in Table 3) as a measure of response instability. We found a small negative correlation of -0.1. If we use the number of cases in which the ratings in both surveys were identical as a measure of response instability, it results in a small positive correlation of 0.09. This indicates, in our opinion, that the order of the item presentation in the survey has only a negligible impact on our results.

Our findings and interpretations rely on the assumption that different ratings of the same item by the same participant in both surveys result from response instability, that is, more or less random memory effects. Of course, such differences can also result from the fact that participants changed their opinions about the product based on recent experiences. We tried to avoid this by limiting the time between the surveys, asking an explicit question about such experiences and removing the corresponding responses, and by offering quite popular products for which we can assume a long usage experience and a high usage frequency. However, it is impossible to completely rule out that some participants did experience some events that changed their impression a bit, but they did not report these events because they didn't consider them fundamentally changing their opinion. Thus, it is important to confirm the results within a different experimental setting.

Tips for Usability Practitioners

- Relatively, there is a lot of response instability in UX surveys. If the same participant judges the same product with the same UX items, results can vary even if the participant has not changed their opinion about the product.
- Response instability is a more or less random effect which is symmetrical (that is, the chances that the scores increase or decrease in secondary evaluations are approximately equal). Thus, it has no effect on the mean scores of items or scales.
- But response instability seems to have a high impact on the observed standard deviations of items and scales. Thus, response instability must be considered in the interpretation of standard deviations. We cannot simply interpret them as a measure of the differences in the opinion of participants concerning the UX aspects measured by the items. A large extent of the variability in the data seems to be caused by response instability or random fluctuations in the responses.

References

Dawes, J., Stocchi, L., & Dall'Olmo-Riley, F. (2020). Over-time variation in individual's customer satisfaction scores. *International Journal of Market Research*, *62*(3), pp. 262–271.

Dickman, S. J., & Meyer, D. E. (1988). Impulsivity and speed-accuracy tradeoffs in information processing. *Journal of Personality and Social Psychology*, *54*(2), pp. 274–290.

Finstadt, K. (2010). The Usability Metric for User Experience. *Interacting with Computers*, *22*(5), pp. 323-327.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), pp. 650–669.

Horst, P. (1966). *Psychological measurement and prediction*. Wadsworth Publishing Company, Inc.

Johnson, E. J., & Payne, J. W. (1985). Effort and accuracy in choice. *Management Science*, *31*(4), pp. 395-414.

Lam, S. S., & Woo, K. S. (1997). Measuring service quality: A test-retest reliability investigation of SERVQUAL. *International Journal of Market Research*, *39*(2), pp. 381–396.

Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In *HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society* [Symposium]. USAB 2008, Graz, Austria, November 20-21, 2008. *Proceedings 4* (pp. 63-76). Springer Berlin Heidelberg.

Lewis, J., & Sauro, J. (2023). *The variability and reliability of standardized UX scales*. MeasuringU. Retrieved January 3, 2024, from https://measuringu.com/reliability-and-variability-of-standardized-ux-scales/

Lewis, J.R., & Erdinc, O. (2017). User Experience rating scales with 7, 11, or 101 points: Does it matter? *Journal of Usability Studies*, *12*(2), pp. 73–91.

Lienert, Gustav A. (1989). *Testaufbau und testanalyse*. Psychologie Verlags Union.

McLellan, S., Muddimer, A., & Peres, S. C. (2012). The effect of experience on system usability scale ratings. *Journal of Usability Studies*, *7*(2), pp. 56–67.

Moshagen, M., & Thielsch, M. T. (2013). A short version of the visual aesthetics of websites inventory. *Behaviour & Information Technology*, *32*(12), pp. 1305–1311.

Reichheld, F. F. (2003). The one number you need to grow. *Harvard Business Review*, *81*(12), pp. 46–55.

Riley, F. D. O., Ehrenberg, A. S. C., Castleberry, S. B., & Barwise, T. P. (1997). The variability of attitudinal repeat-rates. *International Journal of Research in Marketing*, *14*(5), pp. 437–450.

Rungie, C., Laurent, G., Riley, F. D. O., Morrison, D. G., & Roy, T. (2005). Measuring and modeling the (limited) reliability of free choice attitude questions. *International Journal of Research in Marketing*, *22*(3), pp. 309–318.

Sauro, J., & Lewis, J. R. (2011). When designing usability questionnaires, does it hurt to be positive? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2011, May 7-12, 2011*. Vancouver, BC, Canada (pp. 2215–2224).

Sauro, J., & Lewis, J. (2023). *How Variable Are UX Rating Scales? Data from 100,000 Responses*. MeasuringU. Retrieved January 3, 2024, from https://measuringu.com/how-variable-are-ux-rating-scales-data-from-100000-responses/

Sauro, J. (2017). *Measuring Usability: From the SUS to the UMUX-Lite*. MeasuringU. Retrieved August 16, 2024, from https://measuringu.com/umux-lite/

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*, pp. 129–138.

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017a). Design and evaluation of a short version of the User Experience Questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence*, *4*(6), pp. 103–108.

Schrepp, M. (2020). On the usage of Cronbach's Alpha to measure reliability of UX scales. *Journal of Usability Studies*, *15*(4), pp. 247–258.

Schrepp, M. (2021). *User Experience questionnaires: How to use questionnaires to measure the user experience of your products?* KDP. ISBN-13: 979-8736459766.

Schrepp, M., Kollmorgen, J., & Thomaschewski, J. (2023). Impact of Item Polarity on the Scales of the User Experience Questionnaire (UEQ). In *Proceedings of the 19th International Conference on Web Information Systems and Technologies*. WEBIST (pp. 15–25). ISBN 978-989-758-672-9; ISSN 2184-3252, SciTePress. DOI: 10.5220/0012159900003584

Schrepp, M. (2023). *An analysis of standard deviations for UEQ scales*. Researchgate. DOI 10.13140/RG.2.2.34969.08809. Retrieved January 3, 2024, from https://www.researchgate.net/publication/370472467_An_analysis_of_standard_deviations_for_UEQ_scales

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, *54*(2), pp. 93–105.

Strack, F., Martin, L. L., & Schwarz, N. (1988). Priming and communication: Social determinants of information use in judgments of life satisfaction. *European Journal of Social Psychology*, *18*(5), pp. 429–442.

Torkzadeh, G., & Doll, W. J. (1991). Test-retest reliability of the end-user computing satisfaction instrument. *Decision Sciences*, *22*(1), pp. 26–37.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press, Cambridge, UK.

Westbrook, R. A. (1980). A rating scale for measuring product/service satisfaction. *Journal of marketing*, *44*(4), pp. 68–72.

Zaller, J., & Feldman, S. (1992). A simple theory of the survey response: Answering questions versus revealing preferences. *American Journal of Political Science*, *36*(3), pp. 579–616.

About the Authors

Martin Schrepp
Dr. Martin Schrepp studied mathematics and psychology at the University of Heidelberg. Since 1994 he has worked as a UX designer and researcher at SAP SE (Germany). His research interests include HCI, UX evaluation methods, statistics, exploratory data analysis, and cognitive sciences.

Jörg Thomaschewski
Dr. Jörg Thomaschewski became full Professor at the University of Applied Sciences Emden/Leer (Germany) in September 2000. His research interests are human-computer interaction and software engineering. Dr. Thomaschewski founded the research group "Agile Software Development and User Experience" at the University of Applied Sciences Emden/Leer in 2009.

Appendix A

Table A. Original German Items

Item	Item Text
UEQ-S item 1	behindernd/unterstützend
UEQ-S item 2	kompliziert/einfach
UEQ-S item 3	ineffizient/effizient
UEQ-S item 4	verwirrend/übersichtlich
UEQ-S item 5	langweilig/spannend
UEQ-S item 6	uninteressant/interessant
UEQ-S item 7	konventionell/originell
UEQ-S item 8	herkömmlich/neuartig
UMUX item 1	Das Produkt ist einfach zu benutzen.
UMUX item 2	Die Fähigkeiten dieses Produkts erfüllen meine Anforderungen.
UMUX item 3	Die Verwendung dieses Produkts ist eine frustrierende Erfahrung.
UMUX item 4	Ich muss zu viel Zeit damit verbringen, Dinge mit diesem System zu korrigieren.
VISAWI-S item 1	Im Layout passt alles zusammen
VISAWI-S item 2	Das Layout ist angenehm vielseitig
VISAWI-S item 3	Die farbliche Gesamtgestaltung wirkt attraktiv
VISAWI-S item 4	Das Layout ist professionell
NPS	Wie wahrscheinlich ist es, dass sie das Produkt einem Freund oder einer Freundin empfehlen werden?

Table B. Reported Usage Frequencies for the Products

Product	NA	On a daily basis	Several times a week	Several times a month	Several times a year
Instagram	1	16	5	1	0
Moodle	0	5	21	0	0
Amazon	1	1	7	12	3
Spotify	0	17	1	2	0
Netflix	0	3	10	11	1
MS Teams	0	5	5	3	1
Overall	2	47	49	29	5

Table C. Reported Usage Experience for the Products

Product	NA	Less than 6 month	Between 6 and 12 month	Between 1 and 3 years	More than 3 years
Instagram	0	0	1	2	20
Moodle	0	4	1	3	18
Amazon	1	0	1	2	20
Spotify	0	0	1	2	17
Netflix	0	0	0	4	21
MS Teams	1	0	1	8	4
Overall	2	4	5	21	100

Table D. Means and Standard Deviations for All Products and UX Questionnaires

Quest.	S.	Value	Insta-gram	Moodle	Netflix	Amazon	Spotify	MS Teams
UMUX	1	Mean	75	71	81	76	77	78
	1	Std. Dev.	17.01	18.52	19.05	20.55	19.05	20.62
	2	Mean	76	73	84	74	82	75
	2	Std. Dev.	13.23	12.32	15.64	22.37	17.38	23.95
VISAWI-S	1	Mean	5.62	4.57	6.00	4.85	5.73	5.43
	1	Std. Dev.	0.89	1.04	0.92	1.29	0.97	1.13
	2	Mean	5.73	4.41	6.01	5.08	5.94	5.34
	2	Std. Dev.	0.63	0.97	0.81	1.00	0.76	1.05
UEQ-S	1	Mean	5.13	4.06	5.12	4.48	5.23	4.49
	1	Std. Dev.	0.71	0.8	0.81	0.75	0.73	0.87
	2	Mean	5.08	4.13	5.6	4.54	5.34	4.57
	2	Std. Dev.	0.67	0.76	0.85	0.81	0.71	0.92
NPS	1	Mean	7.57	6.27	8.92	8.50	9.45	7.93
	1	Std. Dev.	2.84	2.52	2.02	2.17	1.99	2.79
	2	Mean	7.21	6.65	8.72	8.38	9.60	8.36
	2	Std. Dev.	2.78	2.04	2.51	2.06	1.98	2.50