Comparison of Three One-Question, Post-Task Usability Questionnaires

Jeff Sauro
Oracle Corporation
Denver, CO USA
jeff@measuringusability.com

Joseph S. Dumas
User Experience Consultant
Yarmouth Port, MA USA
joe.dumas99@gmail.com

ABSTRACT
Post-task ratings of difficulty in a usability test have the potential to provide diagnostic information and be an additional measure of user satisfaction. But the ratings need to be reliable as well as easy to use for both respondents and researchers. Three one-question rating types were compared in a study with 26 participants who attempted the same five tasks with two software applications. The types were a Likert scale, a Usability Magnitude Estimation (UME) judgment, and a Subjective Mental Effort Question (SMEQ). All three types could distinguish between the applications with 26 participants, but the Likert and SMEQ types were more sensitive with small sample sizes. Both the Likert and SMEQ types were easy to learn and quick to execute. The online version of the SMEQ question was highly correlated with other measures and had equal sensitivity to the Likert question type.

Author Keywords
Usability evaluation, satisfaction measures, post-task ratings, sensitivity, external validity.

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): User Interfaces–Evaluation/Methodology.

INTRODUCTION
One of the measures of user satisfaction in a usability test is a post-task questionnaire or rating. Among the advantages of this measure are that it can:

- provide diagnostic information about usability issues
- measure user satisfaction immediately after the event, usually the completion of a task, potentially increasing its validity.

It does not measure overall satisfaction with a product, which is the domain of post-test questionnaires, such as the Software Usability Scale (SUS) [1].

Because of the time constraints imposed by a measure made after each task, researchers have worked to make post-task questionnaires brief and easy for participants to use. One of the first post-task questionnaires, the After-Scenario Questionnaire (ASQ), is composed of three rating scales in a Likert[1] format [3]. To assess the impact of the questionnaire, participants performed the same tasks with three software products and filled out the questionnaire after each task. The ASQ exhibited acceptable reliability and sensitivity and the Likert format was easy for participants to use and easy for researchers to score.

A more recent study compared four variations of the Likert question type, including two of the ASQ questions [8]. In that study, each participant used one of the formats to rate tasks. All of the scales had significant correlations with task time and a post-test SUS questionnaire. The analysis also correlated the rating types with the total set of ratings from over 1100 respondents. The format with the highest correlation was similar to that shown in Figure 1 (This versions shows 7 levels whereas [8] had 5.). That format was also the most sensitive at smaller sample sizes.

That study also included a version of Usability Magnitude Estimation (UME) [5]. With UME, users create their own scale. They assign a task rating with any value greater than zero, whether they are rating task difficulty or any other subjective dimension. The judgment and therefore the rating is supposed to be on the basis of ratios. So if Task 1

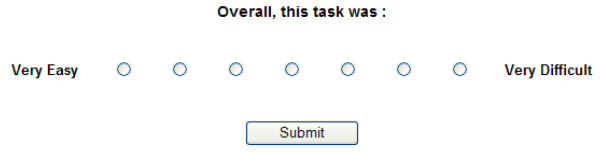

Figure 1. A variant of the Likert scale found most reliable by [8]

CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA.

[1] We recognize that the term "Likert scale" is sometimes used to mean a scale with the exact format that Rensis Likert used and sometimes used to refer to any semantic distance rating. We are using the term in the latter sense.

is rated a 10 and Task 2 is judged twice as difficult, it should be given a rating of 20. The resulting ratings are then converted into a ratio scale of the subjective dimension..

UME was created to overcome some of the disadvantages of Likert scales. Their closed-ended nature may restrict the range of ratings available to respondents (so called ceiling and floor effects) and the resulting scale has questionable interval properties. But the conversion of the raw ratings is achieved with a mathematical formula, which makes UME more burdensome for some researchers than the Likert format.

A UME condition was included in [8] but, in the pilot test, the moderator had difficulty explaining to participants how to make ratio judgments. That study had a remote asynchronous design, which meant that there would be no moderator-participant interaction when participants made their judgments. To simplify the rating, that study constrained the UME scale to values between 1 and 100, making the scale closed ended. The study found that participants treated their version of UME like a Likert scale, ignoring or not understanding how to make ratio judgments.

Our organization conducts many user studies each year and we would like at least some of them to include a post-task satisfaction measure. The studies reported here were attempts to gather some solid data about the value of that measure.

In Experiment 1, we conducted a small-scale study to compare the Likert and UME formats without restricting the range of the UME scale.

In Experiment 2, we added another rating method that has been found to be easy to use: the Subjective Mental Effort Questionnaire (SMEQ) also referred to as the Rating Scale for Mental Effort [10,11]. It consists of a single scale with nine labels from "Not at all hard to do" to "Tremendously hard to do" (See Figure 2.).

In the paper version, participants draw a line through a vertical scale to indicate how much mental effort they had to invest to execute a task. The item positions in a paper format are shown as millimeters above a baseline and the line of the scale runs from 0 to 150, thus leaving quite a large distance above "Tremendously hard to do," which is sometimes used by participants. Scoring the paper version of SMEQ requires measuring the distance in millimeters from the nearest vertical line marking.

In previous studies, SMEQ has been shown to be reliable and easy for participants to use [2,10]. It shares some qualities with the UME in that its upper bound is very open ended. But it may be easier for users to learn than a UME judgment. What's more, the scaled labels (originally in Dutch) were chosen based on psychometrically calibrating them against tasks.

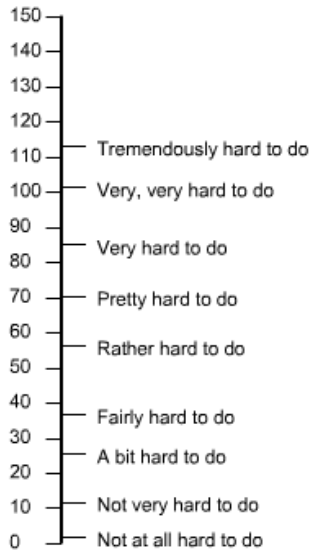

Figure 2. The SMEQ.

Scale Steps

One criticism of the ubiquitous Likert scales is the small number of discrete levels (usually 5 or 7). This limitation could introduce error as respondents are forced into a few choices. The theoretical advantage of UME or SMEQ is their continuous (or at least, near-continuous) number of response choices. With more choices user sentiments can be more precisely recorded.

The more scale steps in a questionnaire item the better, but with rapidly diminishing returns. As the number of scale steps increases from 2 to 20, there is an initial rapid increase in reliability, but it tends to level off at about 7 steps. After 11 steps there is little gain in reliability from increasing the number of steps [6]. The number of steps is important for single-item assessments, but is usually less important when summing scores over a number of items. Attitude scales tend to be highly reliable because the items typically correlate rather highly with one another [4]. Since we are working with single item scales the number of scales steps is important.

The analysis by Tedesco and Tullis [8] showed that a single question is as good as or better than multiple questions in gathering post-task subjective satisfaction. In this study we intended to determine if the gains from more continuous scales outweigh the additional burden compared to the simpler and more familiar Likert scale.

EXPERIMENT 1

In the only study we are aware of [8], a closed-ended magnitude estimation was compared against more traditional Likert-type scales. Prior to the publication of [8] we conducted a small scale study to asses the administrative overhead in using UME as well as to investigate any advantage of UME over the more familiar Likert scales. In

both question types, a higher rating meant the task was more difficult.

Method

Six users attempted seven tasks on a web-based Supply Chain Management application. The participants were experienced with both the domain and an earlier version of the application. Prior to attempting tasks, participants received practice on making ratio judgments by rating the size of circles, the training method used in [5]. Following each of the seven tasks, participants responded to both a UME question and two 7 point Likert scales (See Figure 3). The presentation order was alternated for each task.

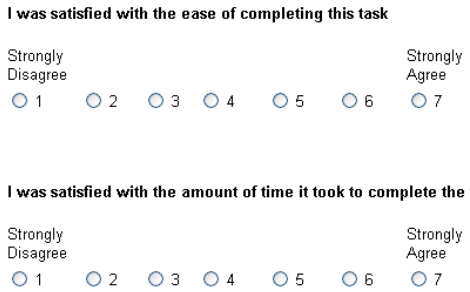

Figure 3: Scales adapted from [3] used in Experiment 1.

Participants spoke their UME rating and the moderator wrote the response down. Participants recoded their responses to the Likert scales in a web-based form.

Results

Scaled UME responses and the average of the two Likert responses correlated very highly (r = 0.84) with each other. Task performance measures were averaged and correlated with each scale. Both measures correlated with task completion rates (UME r = 0.45, Likert r = .5), and errors (UME r = 0.78, Likert r = .73) but not with task time. Only the correlations with errors were statistically significant ($p < .05$) due, we believe, to very low power of this test. The results of a 1-Way ANOVA with UME as the dependent variable found significant differences between tasks ($F(3,44)=4.83$ $p < 0.01$) . Differences between tasks using the Likert scales were not significantly different ($F(3,44)=1.54$ $p > 0.20$).

We found that users had some difficulty, especially early in the sessions, in grasping the ratio judgments. We were also concerned with the potential for bias caused from participants having to say their UME rating out loud.

EXPERIMENT 2

Although the results of Experiment 1 suggested that an open-ended UME has some promise, participants still found it difficult to learn to make ratio judgments. Also, asking participants to make both Likert and UME judgments after each task may have confused them about either question type. Furthermore, the UME training on circle size may not be relevant for subsequent tasks in which the difficulty of using software is being judged.

In Experiment 2, we substantially increased our sample size, added more relevant training and practice on UME, included an online SMEQ, and asked participants to perform only one rating after each task.

Method

There were 26 participants who regularly submit reports for travel and expenses and were experienced computer users. The products were two released versions of a similar travel and expense reporting application allowing users to perform the same five tasks. Ten of the participants had never used either of the applications, while 16 of them had used both. We wanted to see if experience with the products would affect task ratings. The tasks were:

- Create and submit for approval an expense report for attending a meeting
- Update, modify, and submit a saved report
- Change the user preference for the originating city for reports
- Find a previously approved report and its amount
- Create and submit for approval a report for expenses for a customer visit

One purpose of the study was to measure the performance and preference of experienced users. Consequently, each participant was shown a slide slow demonstration of how to perform each task. They then attempted the task. The participants were not asked to think out loud. They were told that we would be recording their task times, but that they should not hurry – rather to work at a steady pace as they would creating reports at work. If they made an error on a task, we asked them to repeat the task immediately. They rated the difficulty of each task using one of the three rating types and then moved on to the slide show for the next task. After completing the five tasks for one application, they moved on to the second application following the same procedure. To minimize carry-over effects we counter-balanced the application order so that half of the participants started with each application. We also counter-balanced the order of the rating types across tasks and products.

After they had completed the tasks for both applications once, we asked participants to complete them again in the same order but without having to view the slide shows. All participants attempted the tasks a second time and gave a rating after each one. If time allowed, we asked them to attempt the tasks and ratings a third time. If participants completed all 30 task attempts, they would provide 10 response ratings for each questionnaire type for each of the 26 participants. The large number of ratings allowed us to address the perceived difficulty of a task and to compare the sensitivity and practical constraints of the rating instruments.

We compared three post-task rating formats:

- The Likert scale shown in Figure 1
- A UME judgment
- An online version of SMEQ

In the UME format, participants assigned a positive number to the task difficulty by entering it into an online form. The more difficult the task, the higher the value. While there is technically no lower bound to UME, in use it can have lower bound limitations if the participant assigns the first task a low value. To avoid this problem, the rating procedure is often created to leave room at the lower end. In this study, we asked participants first to perform a very simple baseline task – to simply click on a search icon (See Figure 4.).

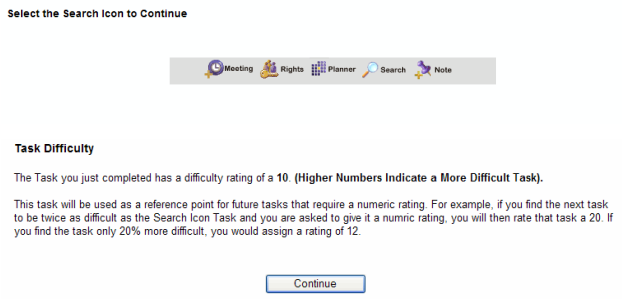

Figure 4. The UME baseline task.

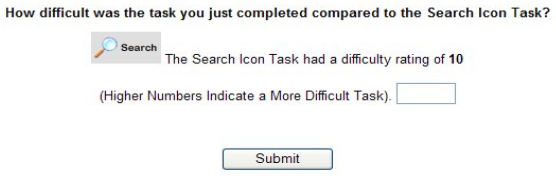

Figure 5. The post-task UME rating.

On subsequent tasks in which the UME format was used, participants were asked to rate task difficulty relative to the simple, baseline task to which we assigned a difficulty of 10. (See Figure 5.)

As noted above, some participants have a difficult time understanding the nature of ratio judgments [8]. Concepts such as "twice as difficult" and "one-half as difficult" take training and feedback to understand. Consequently, we gave participants two practice exercises on UME. Both exercises required making a difficulty rating of a software use task, one easy and one more difficult. We expected that these tasks would provide more relevant practice than rating the size of circles. During the practice, the moderator explained the meaning of their rating to participants. For example, if the participant gave a practice task a rating of 20, the moderator would say, "That means the task was twice as difficult as the baseline search task you did earlier."

For the SMEQ, we created the online version shown in Figure 6 (available at www.usablesurveys.com). In its paper version, the vertical scale is standardized at 15 centimeters high, filling most of a printed page. In our online version, we made each millimeter equal to 2.22 pixels resulting in a scale large enough to fill most of the browser widow on a 1224x768 pixel resolution monitor. Participants moved the slider with a mouse to the point in the scale that represented their judgment of difficulty. The slider "widget" provided the researcher with the scale value to the thousandth decimal. As with all three question types, a more difficult task should be given a higher value.

At the end of the session, participants were asked to complete the ten-item SUS for each application.

Results

There were no significant differences between the sample of ten users with no experience with either product and the 16 who had experience with both. Consequently, we have combined them into one group of 26 for all the analyses shown in this paper.

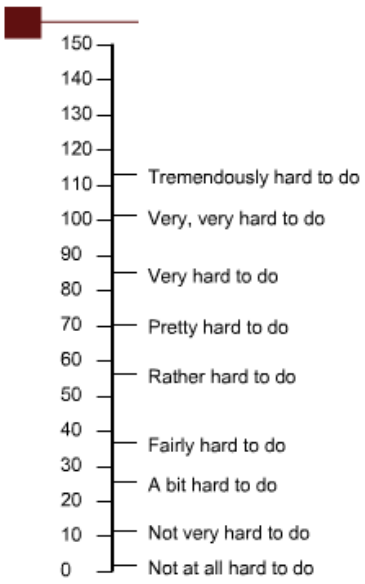

Figure 6. The online SMEQ.

Task	Product A	Product B	Diff.	n
1.	157 (24)	105 (14)	52	16
2.	81 (19)	54 (9)	26	13
3.	52 (13)	34 (6)	18	15
4.	38 (10)	33 (11)	5	18
5.	123 (19)	61 (14)	62	15
Ave	105 (14)	105 (14)	32	15

Table 1. Mean task times (standard deviations) in seconds for the third errorless trial for two products.

Overall, Product A was more difficult to use than Product B. Table 1 shows the average task times for the two products for the third errorless trial. The sample sizes are different because not all of the participants completed all three trials on both products.

All of the mean differences were significant by t-test ($p<.01$) except for Task 4, which was not significant. Figure 7 shows that the post-test SUS ratings also favored Product B.

Diagnostic Value of Post-Task Rating Scales

Figure 7 shows the mean SUS scores for both products. Higher values indicated higher perceived usability. The large gap between the two means makes a very strong case that Product A is perceived as less usable than B. But if one then hopes to improve the usability of Product A, the SUS scores provide little help in identifying which features and functions are contributing most to the poor ratings. It is unclear from the scores alone what is causing this difference in perception.

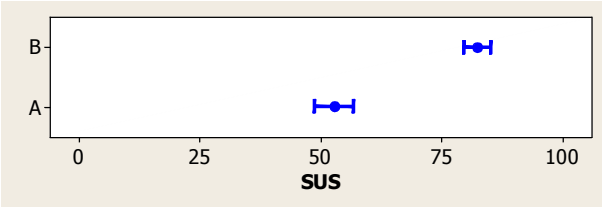

Figure 7. Mean and 95% CI for System Usability Scores for Products B and A (n=26).

In examining the perceived difficulty of individual tasks, Figures 8a-c show the differences in responses by questionnaire type. When the lower boundary of the confidence interval is above the 0 point (meaning no difference), the observed difference is greater than chance (statistically significant). The graphs are oriented so a positive difference indicates a higher perceived difficulty for Product A. Even the least sensitive of the scales, UME (see Figure 8b), discriminated among tasks. Tasks 1 and 2 show a larger gap from the 0 boundary versus less difference between products with respect to the actions and functions encountered for Tasks 3, 4, and 5.

Figures 8a and 8c show the better discrimination for SMEQ and Likert respectively as only Task 4 showed no significant difference in perceived difficulty (T4 crosses the 0 threshold). Tasks 1 and 2 appear to be the largest drivers of the lower perceived usability between products. If one needed a prioritized list of issues in hopes of improving Product A, then Tasks 1 and 2 would be the first ones to examine.

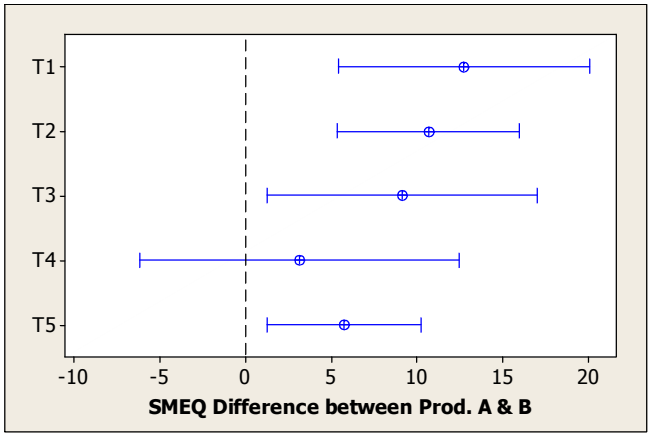

Figure 8a. Differences between products by task and 95% CI for SMEQ. Dashed line shows the 0 boundary, meaning no difference in difficulty between products for that task (T4).

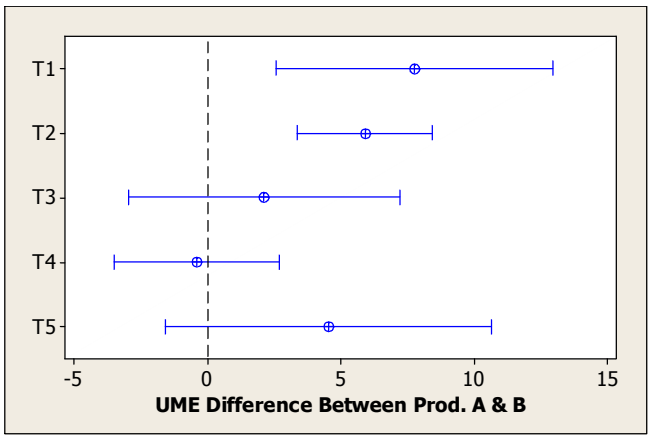

Figure 8b. Differences between products by task and 95% CI for UME. Dashed line shows the 0 boundary, meaning no difference in difficulty between products for that task (T3,T4,T5).

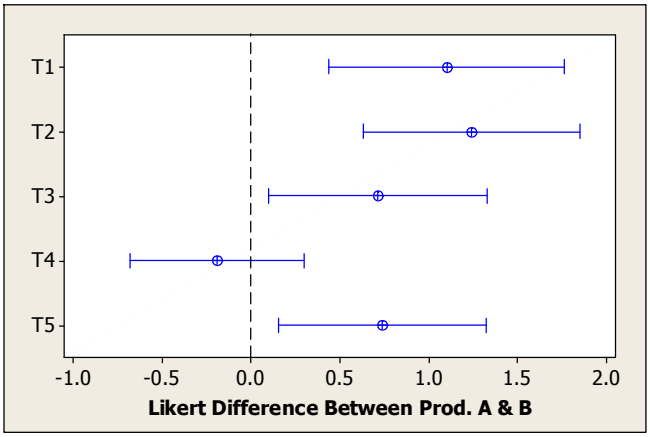

Figure 8c. Differences between products by task and 95% CI for Likert. Dashed line shows the 0 boundary, meaning no difference in difficulty between products for that task (T4).

Sensitivity Resample

To assess the sensitivity of each question type, we took 1000 random samples with replacement at sample sizes of 3, 5, 8, 10, 12, 15, 17, 19, and 20 and compared the means for the two products using a paired-t-test. We used the number of means that were significantly different ($p < .05$) favoring Product B tasks as the dependent variable and sample size, task, and questionnaire as the three independent variables. The more sensitive a questionnaire type is, the more readily it can detect significant differences between products with smaller sample sizes.

The results are displayed in Figures 9 through 12. The results of a 2-Way ANOVA (Question Type, Sample Size) showed a significant difference among questionnaires ($p < 0.05$). A post-hoc comparison test using the Bonferonni correction suggested the only significant difference was between SMEQ and UME ($p < .05$). Figure 9 graphically confirms that SMEQ had higher sensitivity than UME but not from Likert and Likert was not significantly more

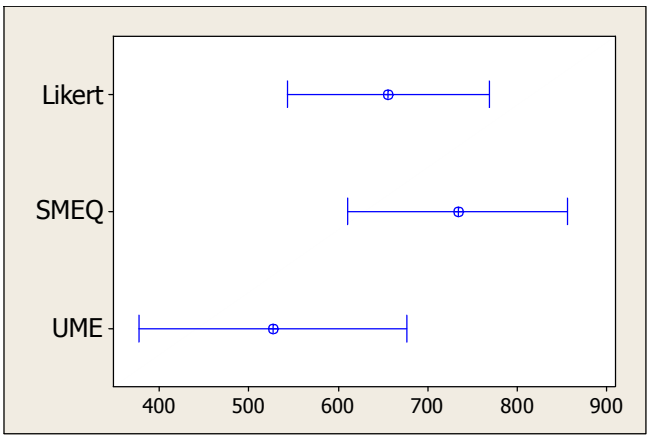

Figure 9. Overall Sensitivity Across Samples Sizes and Tasks. Graph shows mean number of samples found significant out of 1000. Error bars are 95% confidence intervals with Bonferonni Correction.

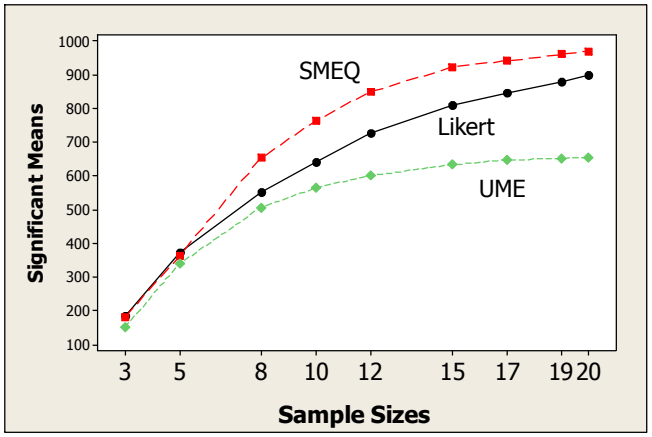

Figure 10. Sensitivity by sample sizes by question type. Y-scale shows the number of significantly different means out of 1000 re-samples.

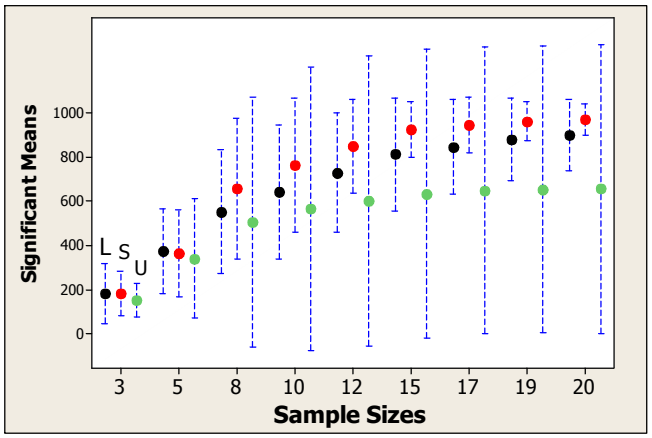

Figure 11. Sensitivity by sample sizes by question type. Graph shows the mean and 95% confidence intervals (L=Likert, S=SMEQ, U=UME).

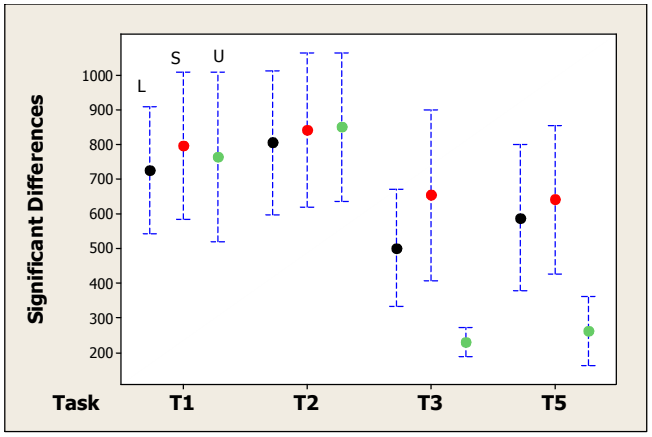

Figure 12. Difference between tasks. Graph shows the mean and 95% confidence intervals (L=Likert, S=SMEQ, U=UME) for the 4 tasks that had a significant difference for the full dataset.

sensitive than UME. Figures 10 and 11 show that UME lags the other two question types across sample sizes larger than 8 and most notably for tasks T3 and T5 (See Figure 12.). There was no significant interaction between sample size and rating type. SMEQ held a slight advantage over Likert, although not large enough to be statistically significant given the sample size.

External Validity

With only one question type per task, we could not evaluate internal consistency using Cronbach's Alpha. We did test how well each rating type correlated with the performance measures taken at the task level (n = 10, *five tasks for two products*). Each participant in this study also filled out the SUS post-test questionnaire for each of the two products. We correlated the mean post-task satisfaction ratings by participant with the SUS score by participant (n = 52, 26 users rating two products). For a discussion on the correlations between performance measures and post-test and post-task ratings scales see [7]. The results of all the correlations are shown in Table 2.

The correlations done at the user summary level (UAO aggregation level—see [7]) showed correlations with SUS of around 0.6 for SMEQ and Likert

	SUS	Time	Comp.	Errors
Likert	-0.568	-.90	.22	-.84*
SMEQ	-0.601	-.82	.88*	-.72
UME	-0.316	-.91	-.05	-.24

Table 2 Correlations (*r*) between rating types and SUS, Task Times, Completion rates, and Errors. Asterisked correlations are significant at the p < .10 all other correlation significant at the p <.01 level.

	Likert	SMEQ	UME
Likert	x		
SMEQ	.940	x	
UME	.955	.845	x

Table 3. Correlations between rating types at the task level. All correlations are significant at the p < .01 level.

and a lower 0.3 for UME (The differences between those correlations were not statistically significant). Both SMEQ and Likert showed statistically higher correlations than UME for completion rates and errors respectively (p <.01). The size and significance of the correlation between SMEQ and task-time is consistent with published data [2,10]. Tests were performed using Fisher-z transformed correlations. Finally we compared the formats against each other. Because only one questionnaire was administered after each task, we could not correlate at the observation level, instead we aggregated the measures at the task level (n = 10) as shown in Table 3 (the TAO aggregation level—see [7]).

In spite of UME's lack of sensitivity, it correlated significantly with the other types as it did in a previous study [8].

SCALE BOUNDARY EFFECTS

The theoretical advantage of UME and SMEQ is their continuous levels. UME has no upper bound. SMEQ has an upper limit but placed substantially above the highest label providing additional upward choices (indicating extremely high mental effort). We wanted to see if participants exploited these advantages across the two products in this study. Since the Likert scale had a maximum of 7 points, we know this is the greatest number of choices we could see expressed by one user across the 10 task rating opportunities. We wanted to know, given a larger spectrum of choices if users would take advantage of them, thus providing evidence that the Likert scale is in fact artificially constraining their judgments.

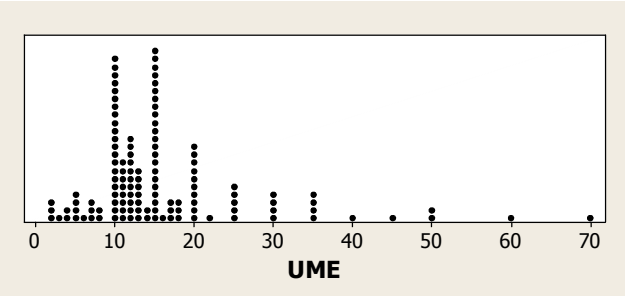

Figure 13. Distribution of UME scores (n = 231) showing a total of 26 distinct scale choices.

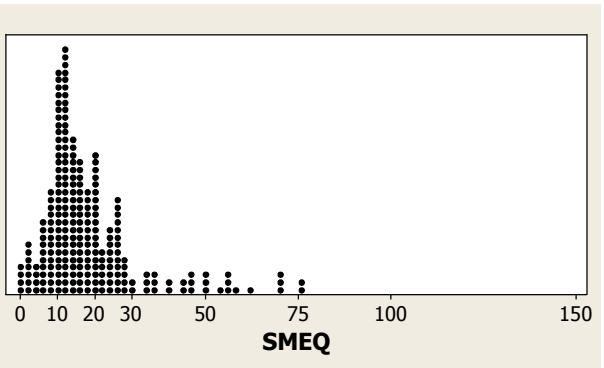

Figure 14. Distribution of SMEQ scores (n = 236) showing a total of 211 distinct scale choices (two SMEQ scale steps per column).

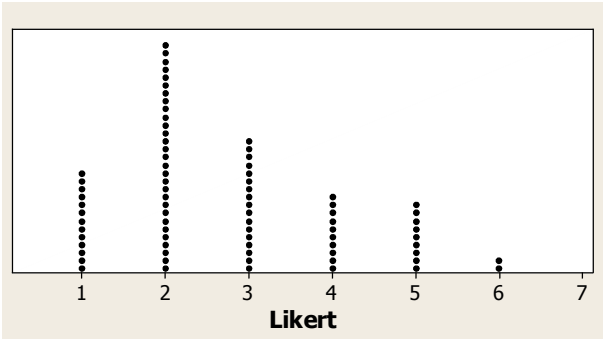

Figure 15. Distribution of Likert scores (n = 234) showing a total of 6 distinct scale choices.

Figures 13, 14, and 15 show the distribution of responses for each question type. Only 6 of the 7 options were selected from the Likert scale while there were 26 and 211 different scale choices used in UME and SMEQ respectively.

Grouping the scale choices by participant, at most there could be 10 distinct choices (1 for each task rating). Figure 16 shows the means and 95% Bonferonni corrected confidence intervals. The mean number of choices participants selected across the 10 tasks for Likert was 3.7 (SD 1.2). For UME the mean number was 5.3 (SD 1.9) and for SMEQ the mean of 9.9 (SD 0.19). For SMEQ only one user selected the same response twice, an unlikely event considering that the slider allowed for responses to the thousandth decimal.

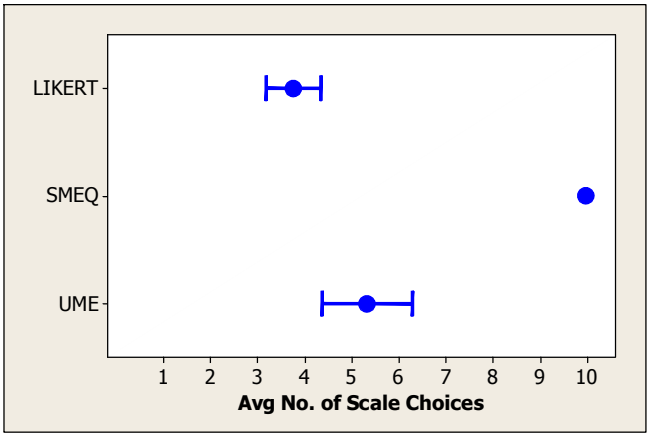

Figure 16. Mean number of choices selected by users across tasks by questionnaire type. Error bars represent the 95% Bonferonni corrected confidence interval

The results of a One-Way ANOVA confirmed the significant difference between means as indicated by the confidence intervals ($F(2,75) = 161.32$; $p < .001$). For UME this data showed that despite having an infinite number of scale steps, the average participant used around 5 choices—more than the 3.7 for the Likert scale but still around half as many as SMEQ.

The nature of the online widget "slider," however, allowed for responses to the thousandth decimal, a degree of accuracy greater than the original 150 millimeter tolerance of the paper-based SMEQ. To reduce the effects of decimal differences between response (e.g. response of say 10.002 and 10.01) we rounded the SMEQ responses to the nearest integer to see how this affected the number of distinct choices. In effect this reduced SMEQ to a 150 point scale like the paper-version. Figures 17 and 18 show the integer data. The mean number of distinct responses by user dropped from 9.9 to 7.9. The differences among questionnaire types still remained statistically different, even with ~20% fewer choices by user for SMEQ.

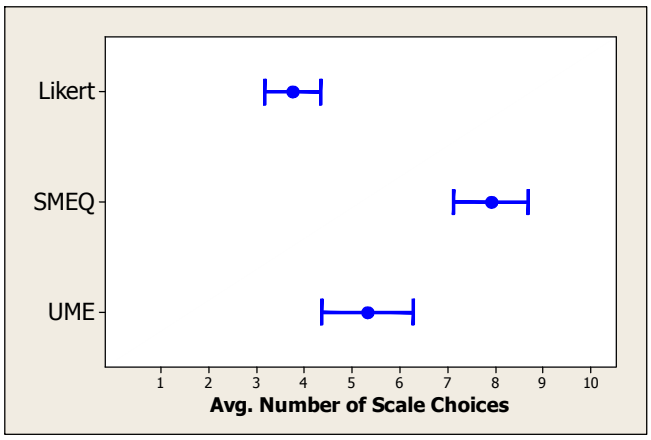

Figure 17. Mean number of distinct choices selected by users across tasks by questionnaire type. SMEQ values are restricted to integers only. Error bars represent the 95% Bonferonni corrected confidence interval

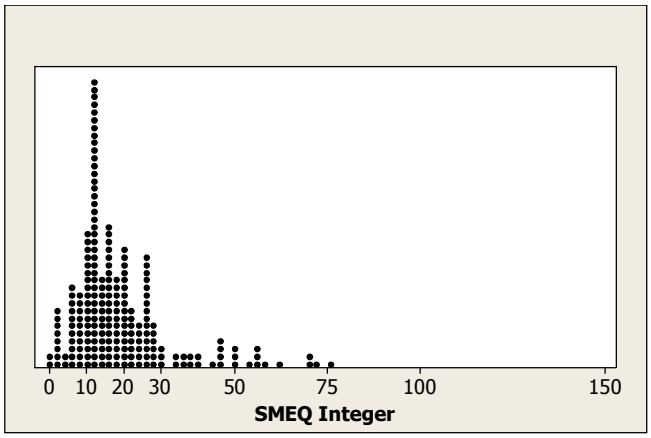

Figure 18. Distribution of SMEQ scores (n = 236) restricted to the nearest integer, showing 48 distinct scale choices (two SMEQ scale steps per column).

DISCUSSION

The task times, error data, and subjective ratings showed that most of the five tested tasks were more difficult for Product A. The high correlations between the post-task ratings and the other measures showed that participants' performance and their perceptions of task difficulty agree.

With a sample size of 26 participants and with hundreds of task ratings, any of the three question types discriminated between many product tasks and correlated highly with both performance and with post-task SUS questions. A tester who used any one of these types would conclude that participants perceived Product A tasks as more difficult to complete. But a closer look shows interesting differences between the three types.

The results of the sensitivity analysis are consistent with a previous study [8]. It is not until participant sample sizes reach the 10-12 level that post-task ratings began to discriminate significantly between product tasks at an 80%

or better level, that is, in this study, 80% of the task means for Product A were significantly higher than for Product B (see Figures 10-12). All of the scales used in both studies showed low sensitivity at the small sample sizes typically used in usability testing. Furthermore, [9] reported a similar sensitivity for post-test questionnaires. It appears that wherever questionnaires and ratings are used during usability test sessions, they may be unable to reliably discriminate between products with large differences in usability at sample sizes below 10.

Both the SMEQ and Likert question types had significantly higher percentages than UME for detecting mean task differences with sample sizes greater than eight. The SMEQ percentages were higher than the Likert percentages, but not significantly so. The UME type asymptotes at about a 60% detection rate with little gain as sample sizes increased from 10 to 20 participants (See Figure 10.).

UME did more poorly than expected in Experiment 2. It had lower sensitivity than the other types and smaller correlations with all of the other measures. The primary reason appears to be participants' inability to understand the concept of ratio judgments in a usability context, which is similar to a previous finding [8]. In spite of providing participants, in Experiment 2, with training, two practice trials rating software use difficulty, establishing a baseline value, and requiring only one rating per task, most of participants' ratings were clustered around the baseline value of 10. While UME has no theoretical upper limit, participants were using it like a closed-ended scale. The anecdotal impression of the moderator was that participants were not making ratio judgments. Because magnitude estimation has worked in other contexts, perhaps we needed to be clearer about what a concept such as "twice as difficult to use" means or provide more training. But those procedures would add more time to test sessions.

The number of different choices used with SMEQ were about twice the number used for UME and three times the number used for the Likert question type. But the SMEQ values still clustered close to the value of 10. We suspect that the baseline value of 10 used for the UME type influenced participants' use of the SMEQ scale. This carryover effect is a price we paid for the higher power gained from using the within-subjects design. Two participants spontaneously said that they used 10 as a baseline for both UME and SMEQ. It is possible that if participants had used only SMEQ, the number of choices used would increase. It would be valuable to repeat this study with a between-subjects design.

The Likert question performed quite well in this study. Participants used it with little or no explanation and it was easy for us to score. It only allowed participants seven levels but the vast majority of participants only used five (and no participant used all 7). Statistically it was as sensitive as SMEQ and had high correlations with all of the other measures. These results justify its popularity as a measure of the difficulty of software use.

The SMEQ question also performed well. Our online version with the slider was easy for participants to use and for us to score, was as sensitive as the Likert question, had high correlations with other measures, and allowed participants to use a larger range of choices. Our anecdotal impression was that participants enjoyed using it more than the other question types. This is the first study we are aware of in which SMEQ has been used in an online format. We do not know if the paper format will yield similar results but we have no reason to believe it would fail to do so.

Finally, it is interesting that these question types were sensitive to task and product differences even when the participants were trained on how to perform the tasks and then performed them for three errorless trials. Perhaps one of the reasons for the high correlations between the post-task question types with the other measures was the fact that all participants were experienced computer users and were repeating the same tasks. However, participants were still making errors and hesitations on their third attempt. Not only do some usability problems persist over repeated trials, but users are sensitive to their presence. Subjective ratings reflect the presence of ease-of-use problems as well as initial ease-of-learning problems.

CONCLUSION

Post-task questions can be a valuable addition to usability tests. They provide additional diagnostic information that post-test questionnaires do not provide and it does not take much time to answer one question. This study and [8] both showed high correlations with other measures, which is evidence of concurrent validity.

With sample sizes of above 10-12, any of the question types used in this study (and in [3]) yield reliable results. But below 10 participants, none of the question types have high detection rates nor do the common post-test questionnaires [9].

The popular Likert question was easy for participants to use, was highly correlated with other measures and, with a seven-point format, did not show a ceiling effect for these tasks and products. What's more, it was easy for test administrators to setup and administer in electronic form.

The SMEQ question showed good overall performance. In its online version, it was easy to learn to use, was highly correlated with other measures, and had equal sensitivity to the Likert question. One draw-back was the requirement of a special web-interface widget. But the widget made the question easy to score.

Participants had difficulty learning to use the UME question type, confirming the findings in [8]. It was less sensitive than the other question types and had lower correlations with other measures. Given the positive results in some

other studies [5] and the long history of magnitude estimation applied to other stimuli, a different training and practice procedure than we used may yield better results.

In addressing our question on whether the benefits of using SMEQ or UME outweigh the additional burden they bring compared to a Likert scale, this analysis suggests the answer is *perhaps* for SMEQ and *probably not* for UME.

REFERENCES

1. Brooke, J. (1996). SUS: A Quick and Dirty Usability Scale. In: P.W. Jordan, B. Thomas, B.A. Weerdmeester & I.L. McClelland (Eds.), U*sability Evaluation in Industry*. London: Taylor & Francis, 189-194.
2. Kirakowski, J. & Cierlik, B. (1998). Measuring the Usability of Web Sites, *Proceedings of the Human factors and Ergonomics Society 42nd Annual Meeting, 424-428.*
3. Lewis, J. R. (1991). Psychometric evaluation of an after-scenario questionnaire for computer usability. studies: The ASQ. *SIGCHI Bulletin, 23, 1, 78-81.*
4. Lewis, J. R. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction, 14, 463–488*.
5. McGee, M. (2004). Master Usability Scaling: Magnitude Estimation and Master Scaling Applied to Usability Measurement. *Proc. CHI 2004* ACM Press (2004), pp. 335-342
6. Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.
7. Sauro, J. & Lewis, J.R.(2009) "Correlations among Prototypical Usability Metrics: Evidence for the Construct of Usability" *Proc. CHI 2009 In Press*.
8. Tedesco, D. & Tullis, T. (2006). A Comparison of Methods for Eliciting Post-Task Subjective Ratings in Usability Testing. *Usability Professionals Association (UPA), 2006, 1-9.*
9. Tullis, T. and Stetson, J. (2004). A Comparison of Questionnaires for Assessing Website Usability. Usability Professionals *Association (UPA), 2004*, 7-11.
10. Zijlstra, F. (1993). *Efficiency in work behavior. A design approach for modern tools.* PhD thesis, Delft University of Technology. Delft, The Netherlands: Delft University Press.
11. Zijlstra, F.R.H & Doorn, L. van (1985). The construction of a scale to measure subjective effort. *Technical Report, Delft University of Technology, Department of Philosophy and Social Sciences.*