# Master Usability Scaling: Magnitude Estimation and Master Scaling Applied to Usability Measurement

**Mick McGee, Ph.D.**
Oracle Corporation
Redwood Shores, California, USA
michael.mcgee@oracle.com

## Abstract
Master Usability Scaling (MUS) is a measurement method for developing a universal usability continuum based on magnitude estimation and master scaling. The universal usability continuum allows true ratio comparisons, potentially between all items measurable by the construct of usability (attributes, tasks, or products -- software or hardware) that have contributed to the meta-set by following the procedures prescribed. This paper describes the background for MUS, data reduction, and cases studies in software usability assessment.

MUS is based on a new measurement method of usability, Usability Magnitude Estimation (UME) [9], where users estimate usability magnitude according to an objective definition of usability. UME allows all items measured within a single usability activity to be compared across one continuum. MUS utilizes UME to assess standard reference tasks across different usability activities to construct one meta-set of data. This meta-set of data can be represented as a universal usability continuum.

MUS is simple to administer, easy to comprehend, and with advanced underlying calculations, powerful to use. The MUS continuum has the potential to be a widespread, robust, universal measurement scale of usability.

## Categories & Subject Descriptors
H.5.2. [**Information Interfaces and Presentation**]: User Interfaces -- *Evaluation/Methodology*.

## General Terms
Experimentation, Human Factors, Measurement.

## Keywords
Usability, master, universal, scale, definition.

## INTRODUCTION
Usability is a multifaceted perceptual phenomenon that is mediated in users by the complex stimuli of a system interface. This research focuses on the ultimate usability measurement in users, in contrast to preliminary assessment by experts [12] or prediction through automated methods [6].

Due to the variety of usability manifestations in users and effectors within system interfaces, usability measurement is difficult to accurately and comprehensively assess, particularly when assessing different systems. Lund, in a call for standardized usability metrics, states that valid and useful usability metrics currently do not exist [8]. In further detail, Frøkjær et al [4], in a survey of CHI published usability studies, showed that the components of usability commonly measured (effectiveness, efficiency, and satisfaction) do not correlate strongly and are not consistently collected. This leads to potentially inaccurate assumptions about overall usability and the risk of ignoring important aspects of usability. They state, "The development of valid and reliable outcome measures is a prerequisite for assessing overall system usability…"

Typical objective measures to assess usability, such as task completion rate, time, errors, and questionnaires, are suitable to intuit areas of an interface needing improvement. They are even frequently cited in summative evaluations for comparing items of interest. However, none can comprehensively assess usability and discriminate among items of interest effectively.

The fundamental problem with task-based performance measures is their susceptibility to the arbitrary makeup of tasks. 'Task' is often described as the basic unit of a usability evaluation; however, task creation is highly dependent on the usability professional. The subjective number of steps chosen and complexity for a given task correlates directly to performance.

Task complexity can cause task completion rates to be so high and errors so low (or vice versa) that no meaningful discrimination can occur. These are the 'ceiling' and 'floor' effects. These effects occur when an overly or under-sensitive measure is used that narrows the range of values available to discriminate conditions [1].

Beyond the confounding relationship with task and sensitivity issues, the ability to generalize results based on performance metrics is poor. Performance metrics provide a narrow assessment of overall multifaceted usability, limited

to only the specific quality they are measuring. Corroborating the Frøkjær et al [4] findings, Oracle usability activities have frequently shown users cite significant usability problems for products where performance metrics are satisfactory.

Subjective measures are typically large questionnaires or Likert scale ratings. Questionnaires used for computing metrics of usability, e.g., SUMI (Software Usability Measurement Inventory) [7] and SUS (System Usability Scale) [3], are limited to overall evaluations and are impractical to administer at the task level. It is possible to give Likert style ratings for each task of a usability study; however, Likert scales have limited pre-defined ranges tending towards a narrow variance in participant responses, leading to limited differentiability.

Validity is also questionable for both large questionnaires and Likert scales when used for computing overall numerical scores. These types of assessment methods are usually based on ordinal scales with assumed underlying continuums. Individual rating scales with endpoints anchored by adjectives (e.g., Consistent and Inconsistent) rarely validate that one unit difference within the defined range is the same as another, the qualification for an interval scale (a minimum prerequisite for parametrical statistical analysis, which are based on an assumed underlying normal distribution). A questionnaire composed of many rating scales multiplies this validity problem further, as multiple scales are highly unlikely to be equivalent (each summing units of unknown size across individual scales). These problems are usually overlooked and, rather than limiting themselves to ordinal conclusions, researchers often make overall statistical comparisons with composite scores regardless. However, common practice does not alleviate these important inherent problems.

### Usability Magnitude Estimation

UME was developed in response to the described deficiencies of traditional usability metrics [9]. Magnitude estimation of psychological phenomenon is a highly documented method that has proven extremely efficient, ideal for a large number stimuli, and superior to ordinal scales [5]. The basic premise behind magnitude estimation is that humans are good integrators of complex stimuli that enable them to provide unified judgments for abstract constructs.

Investigators have successfully applied magnitude estimation to many multifaceted psychological perceptions mediated by complex phenomenon; e.g., trial evidence (physical stimulus) with guilt (perception) [5]; life events with emotional stress [5]; psychotic symptoms with severity diagnosis of mental disorder [5]; environmental conditions and odor [2]; virtual environments and presence [14]; and virtual environments and cybersickness [10]. Given the multifaceted causes and manifestations of user-perceived usability and similar past applications, use of the magnitude estimation methodology for measuring usability seemed a natural fit.

Thus, UME is a subjective measure of usability based on a users' perception of usability. In practice, users base their perception of usability on an objective definition provided by the usability engineer. Unlike other subjective usability measures, which are usually classified as 'satisfaction' metrics, a broadly defined objective definition allows all aspects of usability to be included; e.g., efficiency and effectiveness in accomplishing a task, as well as satisfaction.

With the objective definition of usability, users are instructed to make ratio estimates, without anchors; i.e., any positive number may be assigned to a target as long as it relates to previous targets. For data reduction, ratio estimates are normalized through a geometric averaging procedure (that preserves ratio information from the raw data) to form a single, ratio scale of usability. The task or target averages taken from the usability scale of a single study are appropriate for statistical analysis.

McGee showed UME to have a number of advantages over traditional measures of usability in efficiency of data collection, sensitivity in detecting differences, robustness to arbitrary task differences, and effectiveness at differentiating isolated tasks or interface elements [9]. However, UME by itself is not appropriate for use across different studies.

A UME scale created in one usability activity cannot be compared to another that tests a different product or set of tasks. Gescheider cites an example of noise measured in a downtown New York City street and a Pennsylvania industrial plant [5]. If the city street averages 50 noise units and the plant 40, the conclusion cannot be drawn that the city street is noisier than the industrial plant. The individual noise scales created are unique to the testing situation. Participants in the New York study may simply use larger number values when making magnitude estimates.

If participants had evaluated reference noises in the New York and Pennsylvania studies, then comparisons could be made. For example, say an 80 dB tone in the New York study received a noise estimate of 100 on the street noise scale and 20 on the Pennsylvania industrial plant noise scale. Now it would be clear that the industrial noise of the Pennsylvania plant is worse than the New York City street. (40/20 > 50/100). The concept of using reference comparisons to compare two different scales based on the same objective definition was formalized by Berglund in a process called master scaling [2].

### Master Scaling

Berglund's master scaling procedure involves a full set of reference items measured separately by an independent set of participants [2]. Subsequent experiments, using magnitude estimation with the same objective definition, then have participants measure conditions not only for that experiments' needs, but also the entire reference scale. The resultant scale in the new experiment is transformed based on the standard reference scale collected in the independent study.

There are several problems with the Berglund master scaling approach, however, when considering a practical measure for usability. First, an independent experiment to construct a standard reference scale would be expensive (Berglund collected data on 30 participants). If this was the only resource concern, the expense could possibly be rationalized.

The second resource problem using this method is the expense in time and money for collecting all the reference estimates for every participant in each usability activity. For nearly all usability activities, this would not be practical or possible within the time and resource constraints of a testing organization.

Lastly, the transformation procedure is likely to alter the original ratio relationships of the target data once transformed by the reference scale. Ratio relationships of the original data are not protected when adjusting all points along one scale to another. Ratio information is the key empirical knowledge gained by magnitude estimation. To sacrifice the validity of the information for the sake of a comparison continuum is questionable.

### Goal and Objectives

The goal of the MUS research was to extend the UME approach to incorporate master scaling. The objective was to: 1) adapt the Berglund approach into a format suitable for usability testing, accounting for the problems identified; and, 2) validate the use of MUS in actual usability activities.

## METHOD

### Data Collection

Collecting MUS data is simple and straightforward. The main part of data collection is exactly the same as UME, with one addition. When collecting UME data in a user session, participants complete a short generic magnitude estimation practice task and estimate magnitude usability after each task or target. MUS adds a third component, estimating the magnitude usability of standardized reference tasks, after the main session estimates have been completed.

To begin the UME data collection process, a generic practice task is performed, before evaluation starts, to ensure the concept of magnitude estimation is understood. The participant reads instructions for magnitude estimation and then rates a set of standardized objects with a known physical scale. Oracle usability activities have successfully used size of circles and length of lines for practice tasks.

Continuing the UME process, just prior to beginning the main evaluation and usually after the practice task, the objective definition of usability is provided. Participants use this definition to estimate magnitude usability after each task. The definition of usability for UME, derived through research in conducting usability activities within Oracle, is,

> "Usability is your perception of how consistent, efficient, productive, organized, easy to use, intuitive, and straightforward it is to accomplish tasks within a system."

This definition is intentionally worded to allow measurement of usability in a variety of scenarios beyond productivity-oriented software, such as Oracle products, by framing the definition with the generic "to accomplish tasks within a system." Individual testing organizations could tailor the definition to their needs and create individual MUS continuums. However, the consequence would be the inability to add and compare data in a meta-set that uses the same definition and reference tasks.

To conclude the MUS addition to the UME process, the standardized reference tasks are evaluated at the end of a usability activity to prevent biasing measurement of the primary experimental conditions. Assessing the reference tasks first would set de facto anchors to the magnitude estimation scale, damaging sensitivity and the ability to differentiate.

### Standardized Reference Tasks

The standardized reference tasks are the basis for combining data from otherwise independent studies. For the purposes of creating a universal usability continuum, it does not matter exactly what the reference tasks are, as long as they are consistently used between studies. They only need to be measurable along the dimension of interest; i.e., usability. As with the definition of usability used, individual organizations could use reference tasks with more face validity to their testing objectives. However, as with altering the usability definition, comparisons with a meta-set would not be possible without normalizing the new reference tasks with known reference values.

To obtain reference ratings in Oracle tests employing MUS, usability engineers direct participants to a web page with the instructions below. To maintain consistency between activities, no other information is provided unless the participants ask specific questions.

> "You will complete two additional tasks using the 'usability ruler' you created previously. Based on your usability scale, provide an estimate of usability for the following two interfaces. The task is the same for both:

*You recently became a member of a website. You used your first name for the username and your last name for the password. Login to the website.*

Note: The websites are not real. They are for test purposes only. No actual login will occur.

1. Website 1 2. Website 2"

The first reference login task, shown in Figure 1, is for a website purposely designed to be simple and clear. The intention is to have a standardized reference with reasonably high usability ratings. There is no intent to obtain the highest possible usability rating (or lowest through website 2), as there is no theoretical limit. Furthermore, user estimates at extreme ends of a continuum can vary widely [1]. The only requirement is a consistent standard reference task.

The second reference login task, shown in Figures 2 and 3, is for a website purposely designed to have low usability: the contrast is reduced (the background is bright green but the text is still readable); the actual login is on the second page, accessed from a rollover link at the bottom of the first page; there are three login fields when only two are required; and, there are three possible buttons to complete the task. This website is expected to receive comparatively low usability ratings.

A second reference is included to average across standard estimates and provide benchmarks to compare future high and low usability experimental conditions. Instead of many reference conditions, as described by Berglund to develop a transformation scale, two are used in MUS to limit the time used within usability activities. The reference tasks shown are also simple, allowing consistent experiences across users. Moreover, with the data reduction methods employed by MUS, only a 'pivot point' is needed, not a full transformation scale.

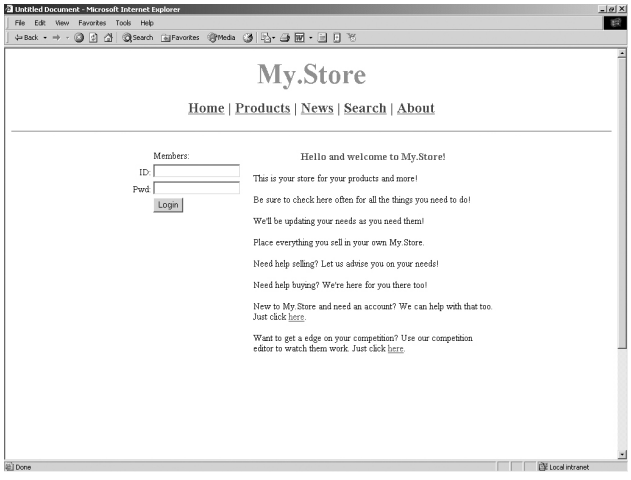


**Figure 1. Reference task website one.**

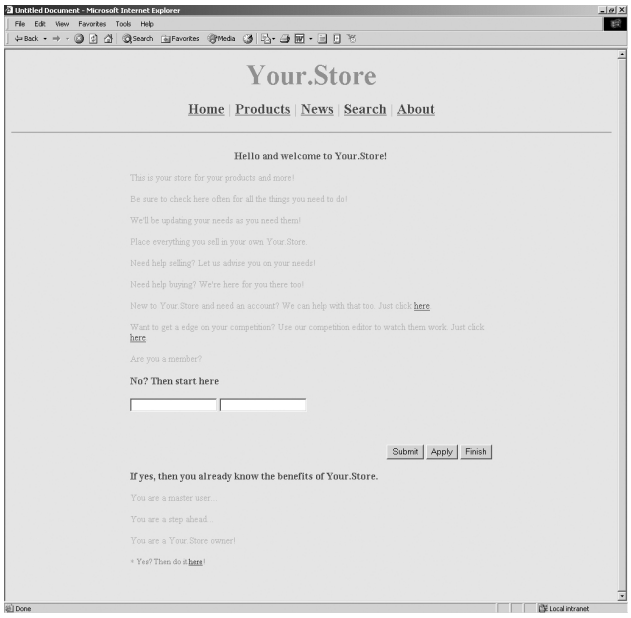


**Figure 2. Reference task website two, page one.**

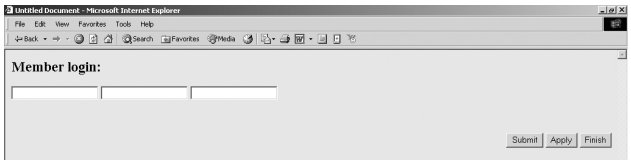


**Figure 3. Reference task website two, page two.**

## Data Reduction

### *Usability Magnitude Estimation*

Data reduction for a single usability activity follows the conventions for magnitude estimation originated by Stevens [15]; described generally by Gescheider [5]; listed step-by-step by Snow [14]; and, shown in table format for UME by McGee [9]. Essentially, a geometric averaging procedure is used on the original data to construct an overall continuum of usability. The fundamental characteristic using geometric averaging provides is the preservation of ratio information obtained through magnitude estimation. This allows any subsequent comparisons to be made along a true ratio scale.

The steps for UME data reduction are:

1. Log scores are calculated for all the original estimates.
2. An overall average score is determined for the entire study.
3. Overall averages are calculated for each participant.
4. All log scores regress to the overall mean by subtracting the difference of each participant overall average (step 3) from the entire study overall average

(step 2) -- (called the 'participant offset'; i.e., 'pivot point').

5. The antilog is taken on all the regressed scores. The resultant continuum, having preserved the ratio information of the original scores, is appropriate for parametric statistical analyses (assuming additional applicable experimental design constraints have been followed).

In formulaic terms, for each estimate:

Normalized Est. = Natural Log (Log (Raw Est.) - Part. Offset)

Where,

Part. Offset = Part. Ave. (Log (Raw Est.)) - Overall Ave. (Log (Raw Est.))

Part. is Participant; Ave. is Average; and, Est. is Estimate.

Table 1 is a UME example from McGee showing preservation of ratio information [9]. To verify, consider the last column in Table 1 (U′), the ratio 17.89 to 4.47 equals 4 to 1; exactly the ratio both participants provided in their initial estimates (column 3 -- U) between tasks four and one.

**Table 1. Example UME data reduction.**

| Participant | Task | U | Log U | Offset | Log U′ | U′ |
|---|---|---|---|---|---|---|
| 1 | 1 | 10 | 1.00 | 0.35 | 0.65 | 4.47 |
| 1 | 2 | 20 | 1.30 | 0.35 | 0.95 | 8.94 |
| 1 | 3 | 30 | 1.48 | 0.35 | 1.13 | 13.42 |
| 1 | 4 | 40 | 1.60 | 0.35 | 1.25 | 17.89 |
| | | | $\bar{x}_1 = 1.35$ | | | |
| 2 | 1 | 2 | 0.30 | -0.35 | 0.65 | 4.47 |
| 2 | 2 | 4 | 0.60 | -0.35 | 0.95 | 8.94 |
| 2 | 3 | 6 | 0.78 | -0.35 | 1.13 | 13.42 |
| 2 | 4 | 8 | 0.90 | -0.35 | 1.25 | 17.89 |
| | | | $\bar{x}_2 = 0.65$ | | | |
| | | | $\bar{x}_{Total} = 1.00$ | | | |

U Raw participant usability scores.
Log U′ Log U minus offset.
U′ Geometrically averaged usability scores; i.e., normalized scores.

### *Master Usability Scaling*

The UME data reduction process preserves the ratio information provided by the original user estimates within a single study. This allows parametric statistical analyses to be used for comparing tasks within that study. However, comparisons between two studies cannot be conducted unless a universal usability continuum is created (see previous industrial noise example). MUS allows the creation of that continuum.

The MUS data reduction process is the same as UME, with one key change. The offsets are computed using the between-study reference task estimates instead of individual within-study estimates. The reference task estimates are the only data values across studies that can be considered the same dataset. Since the reference tasks themselves are from different tests, these reference offsets must first be computed with geometric averaging to preserve their ratio information (assuming more than one reference task is tested across users). Then, a pseudo UME geometric averaging can occur with the substituted offsets. This will normalize all the individual estimates, for all the studies included in the meta-set, by the normalized average reference estimates per user. Thus, MUS essentially requires two rounds of geometric averaging, one for finding the reference offsets, and a second to create the overall continuum.

The steps for MUS data reduction are:

1. Log scores are calculated for all the *reference* estimates for all the studies (and only the reference estimates).
2. An overall reference average score is determined across all the studies.
3. Overall reference averages are calculated for each participant in each study.
4. All log reference scores regress to the overall reference mean by subtracting the difference of each participant overall reference average (step 3) from the entire meta-set overall reference average (step 2) -- (called the 'participant *reference* offset').
5. Then, the UME geometric averaging data reduction process is used across all the data with the 'participant *reference* offsets' substituting for the 'participant offsets' (thus, steps 2 & 3 shown in the UME data reduction are not necessary, and step 4 uses the substituted offsets).

### MUS Universal Usability Continuum

The output of MUS data reduction is a meta-set of data forming a universal usability continuum (labeled U"). Traditional statistical analyses are not appropriate with this continuum, as it is with UME within a single study, since experimental conditions between studies are not controlled.

To make comparisons, specific values of interest can be examined on the MUS universal continuum in their true ratio form. Figure 4 shows the theoretical usability continuum, plotting MUS usability against itself, where an infinitesimally small positive number near 0 is the lowest possible U" value, and infinity the highest. Individual items can be highlighted on the MUS continuum by plotting the average usability scores of selected variables of interest (example shown in Validation).

True ratio comparisons make strong, valid arguments for determining relative usability. However, no statistic yet exists for comparing items of interest across the MUS continuum. Each comparison must be judged on the size of the difference, the number of users represented by the data, and previous investigations employing UME or MUS.

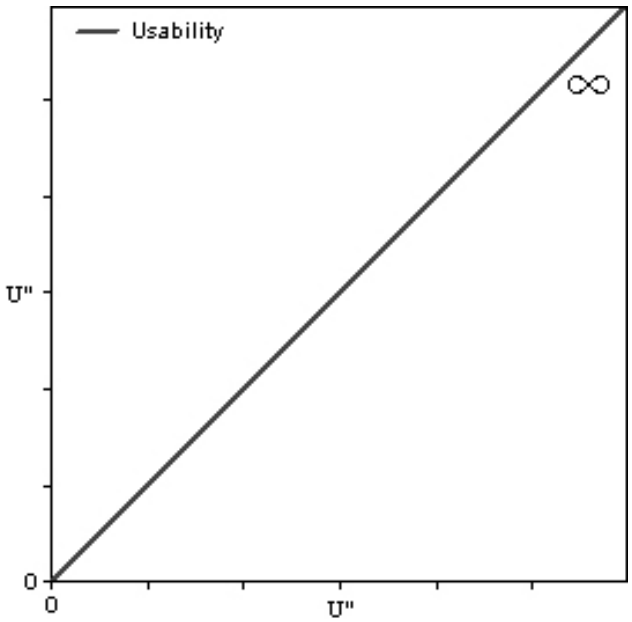


**Figure 4. Theoretical MUS continuum.**

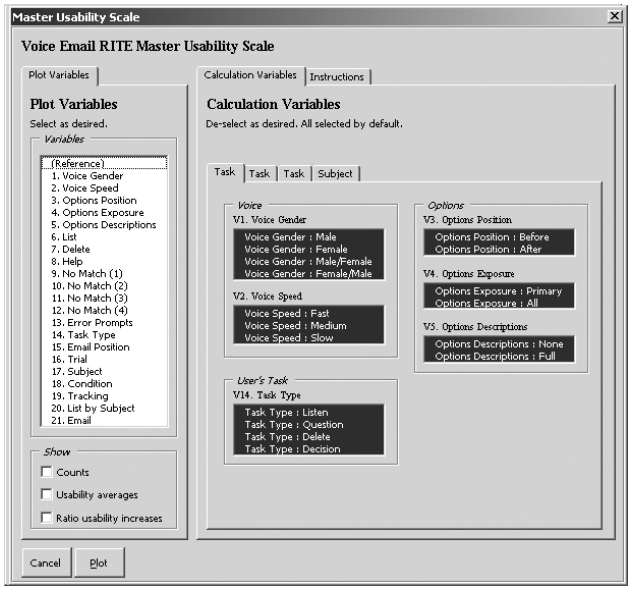


**Figure 5. IVR email evaluation variable selection tool.**

## VALIDATION

### Case Study 1: Formative Evaluation of IVR Email

Oracle has completed two usability activities utilizing MUS. The first was a formative Rapid Iterative Test and Evaluation (RITE) [11], for rapidly improving design through user testing, on a prototype Interactive Voice Response (IVR) email system. The fundamental element of the RITE methodology is allowing designs to change after each user. This variability in design between users, essentially a series of mini 1-user usability tests, posed a significant challenge to measuring usability and justifying design improvements. The considerations for making design decisions were threefold: user feedback, designer expertise, and MUS for the respective tasks and designs.

User feedback and designer expertise are subject to interpretation and bias. To facilitate decision-making where consensus could not be reached, MUS was used as an objective arbitrator by examining the magnitude of difference between designs to determine if changes were necessary. Medlock et al used errors for this assessment [11]; however, errors were too few in our evaluation to indicate any differentiation among designs. We encountered similar problems with other traditional metrics as discussed previously. We utilized MUS instead of UME since the design stimuli were different for each participant. Each participant in essence provided a separate set of UME data needing to be combined into one meta-set.

#### *Results*

Upon concatenation of all the data into one meta-set through MUS, the main challenge was viewing only the relevant variables of interest for particular design variables. There were literally millions of possible combinations that could be displayed on the continuum across the 21 design variables tested. To overcome this issue for the IVR user evaluation, a tool, shown in Figure 5, was constructed with Microsoft Excel Visual Basic to select variables for display on the MUS continuum (the tool is shown for example purposes only, a variety of methods can be employed to examine specific continuum data). The tool incorporated main variable and interaction display capabilities.

Main variables could be selected with the left side of the tool, Plot Variables, to show mean usability scores for all levels of each variable selected, collapsed across all other variables. For example, the variable means associated with recognition errors ("No Match"), where the system failed to recognize user commands, are shown in Figure 6. (Note: The linear continuum was shown in this figure to maintain continuity from the theoretical representation shown previously. A bar graph can also be used to represent the full ratio differences more faithfully.) The resultant MUS continuum shows there are no substantial differences between no recognition errors, one error, or two consecutive errors; and a substantial decrease in usability, 17-20%, to the 3rd and 4th consecutive errors. This proved to be valuable information for designing error prompts for different recognition errors.

The variable selection tool also allowed interactions to be computed by limiting MUS data to certain variable levels (selected through the Calculation Variables) and calculating the respective MUS continuums. Plotting interactions on a

single continuum can provide some information on spread;

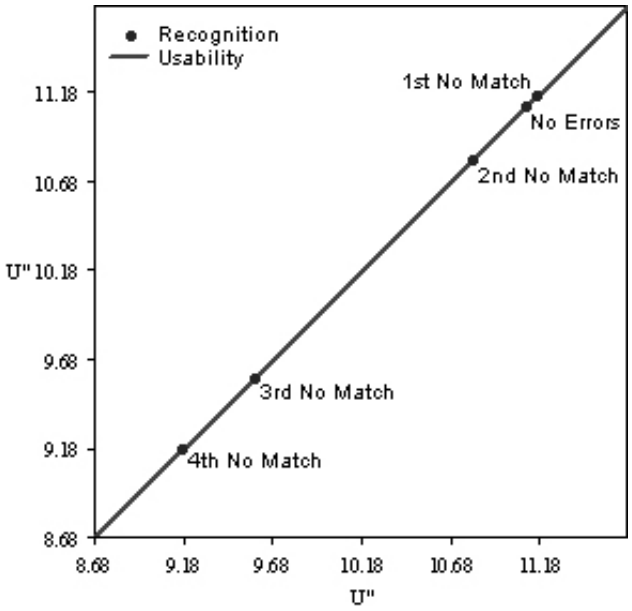


**Figure 6. MUS for recognition error variable levels.**

however, they are best viewed in a traditional interaction graph. These graphs are easily created by plotting main variable levels along the x-axis with the measure, U", on the y-axis. In essence, each main variable level along the x-axis is an individual MUS continuum with data limited to that level. For example, the interaction between recognition errors and error prompt type is shown in Figure 7. The 3rd and 4th consecutive errors, identified in the MUS continuum of Figure 6 as problematic, can now be seen to be more of an issue with repetitive prompts than progressive prompts. This information further helped refine our designs for error prompts in IVR email.

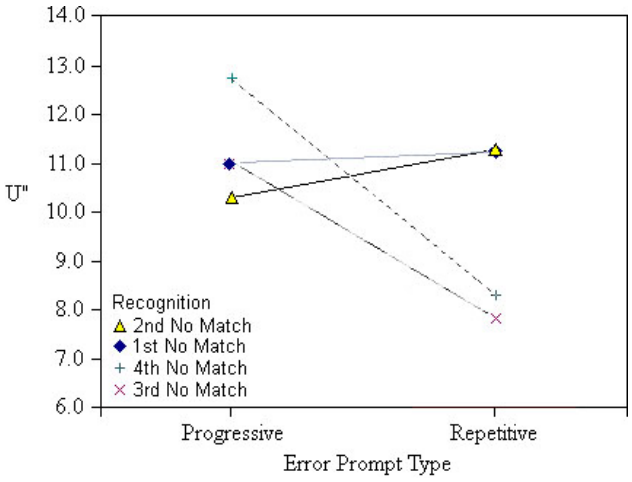


**Figure 7. Interaction example.**

MUS in this RITE study, where designs changed after each user, proved beneficial for confirming design decisions based on user feedback and designer expertise, and essential for reaching decisions in situations where consensus could not be reached. Objective comparisons between the consecutive '1-user studies' would not have been possible without it.

### Case Study 2: Summative Evaluation of PDA Office Applications

The second usability activity utilizing MUS was a formal benchmark test on PDA office collaboration applications (e.g., email, directory, address book). This test followed conventions of the Common Industry Format (CIF) for usability testing and reporting [13]. To incorporate MUS into the analysis of this test, users made magnitude estimates of usability after each task and rated the MUS reference tasks at the end of the test as described previously.

The design for this test was straightforward. All users performed the same 14 tasks on the same product and evaluated usability accordingly. Analysis did not require MUS within this test; however, the product teams had a desire to know how developing products compared against mature products. For this comparison, we used MUS to contrast the prototype voice email usability with the PDA office collaboration applications, both part of the same wireless applications suite.

To simplify results for the development team, the overall usability means for the two products were presented in relation to reference task one, the 'good' usability task. Assigning a value of 100% to reference task one yielded the table of results shown in Table 2.

**Table 2. % Usability.**

| Item | U" / Ref. Task 1 U" (%) |
|---|---|
| Reference Task 1 | 100.00% |
| Mean PDA Email | 69.49% |
| Mean Voice Apps | 57.33% |

The comparison showed that voice applications still have considerable room to improve, and that the production quality PDA applications are better, but still significantly less usable than a simple desktop web task. Most importantly, a benchmark to make valid, true ratio comparisons to all future products utilizing MUS is in place.

## CONCLUSIONS

MUS is a flexible, powerful tool for usability practitioners that meets the challenge set forth by Lund [8] and addresses the problems in current practice as identified by Frøkjær et al [4]. The usability measured is comprehensive, differentiable at the task level, robust to task confounding in overall averages, simple and practical to collect, sensitive in

a variety of experimental conditions, and a valid, true ratio scale continuum across all studies included in the meta-set.

Usability practitioners can use the methodology to assess interface attributes, tasks, series of interactions, whole products, or any other combination of 'usability targets'. As long as the construct of usability can be evaluated, MUS is a potential tool of measurement that can be as sophisticated as needed and still easy to understand. When the methods prescribed are used, MUS can provide a universal usability continuum.

## FUTURE WORK

The most important future work concerning MUS is continuing to validate the MUS procedures and universal usability continuum with more studies. Validating the procedures themselves increases the validity of the method. Including more data in the meta-set increases the robustness of the comparisons.

With enough validation data, two additional statistical advances could be made. New additions to the meta-set would only slightly affect the underlying MUS offsets. In time, the continuum would be very stable, allowing a "true" usability unit to be defined. Furthermore, a statistic could be developed to help determine the extent of the difference between one item and another.

To facilitate future MUS work, the tools to analyze UME and MUS must be easier to use. The analysis tool should require only raw data input without any visible knowledge of the geometric averaging. The selection and display tool should be simple and clear in selecting variables and/or interactions to display. The current procedures as described are reproducible by any usability practitioner; however, they are sophisticated enough to pose a significant barrier to widespread MUS adoption. Work is ongoing to improve both of these analysis tools.

The MUS methodology was purposely developed to potentially apply to any scenario where the construct of usability could be measured. The most logical extension beyond the software user interface validation presented in this paper is in hardware user interfaces. The theoretical limits of the MUS universal continuum are constrained only by the broadly defined operational definition of usability and the practical necessities of collecting information from users.

## ACKNOWLEDGMENTS

The author thanks Oracle Corporation, specifically the Usability and User Interface Design group, for providing the facilities and opportunity to conduct research within usability testing, and Joe Dumas and Aaron Rich for commenting on previous versions of this paper.

## REFERENCES

[1] Badia, P. and Runyon, R.P. (1982). *Fundamentals of behavioral research*. Random House.

[2] Berglund, M.B. (1991). Quality assurance in environmental psychophysics. In S.J. Bolanowski, & G.A. Gescheider, (Eds.), *Ratio scaling of psychological magnitude: In honor of the memory of S.S. Stevens.* Lawrence Erlbaum Associates, Publishers.

[3] Brooke, J. (1996). A "quick and dirty" usability scale. In Jordan, P., Thomas, B., and Weerdmeester, B. (Eds.), *Usability Evaluation in Industry*. UK: Taylor and Francis.

[4] Frøkjær, E., Hertzum, M., and Hornbæk, K. (2000). Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? *Proc. CHI 2000*, 345-352.

[5] Gescheider, G.A. (1997). *Psychophysics: The Fundamentals, 3rd Ed*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

[6] Ivory, M. and Hearst, M. (2001). The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys*, 33(4), 470-516.

[7] Kirakowski, J. (1996). The software usability measurement inventory: Background and usage. In Jordan, P., Thomas, B., and Weerdmeester, B. (Eds.), *Usability Evaluation in Industry*. UK: Taylor and Francis.

[8] Lund, A. (1998). The need for a standardized set of usability metrics. *Proc. HFES, 42nd Annual Meeting*, 688-691.

[9] McGee, M. (2003). Usability magnitude estimation. *Proc. HFES, 47th Annual Meeting*, (691-695).

[10] McGee, M. (1998). Assessing negative side effects in virtual environments. Masters Thesis. Virginia Tech.

[11] Medlock, M.C., Wixon, D., Terrano, M., Romero, R.L., and Fulton, B. (2002). Using the RITE method to improve products: A definition and a case study. *Proc. Usability Professionals Association*.

[12] Nielsen, J. (1993). *Usability engineering*. Academic Press.

[13] NIST. (2001). *Common industry format for usability reports*. National Institute of Standards and Technology.

[14] Snow, M.P. and Williges, R.C. (1998). Empirical models based on free-modulus magnitude estimation of perceived presence in virtual environments. *Human Factors*, 40(3), 386-402.

[15] Stevens, S.S. (1971). Issues in psychophysical measurement. *Psychological Review*, 78(5), 426-45.