Usability Magnitude Estimation

Mick McGee, Ph.D.
Usability & Interface Design
Oracle Corporation
Redwood Shores, California

Magnitude estimation is a measurement method that is very useful for measuring multi-faceted constructs that do not have a physical analog (e.g., usability) and are produced from multidimensional stimuli (e.g., user interfaces). Traditional usability metrics have shown limitations in usefulness and validity. Usability magnitude estimation is an assessment method whereby participants assign usability 'values' to tasks, conditions, or other user interface targets according to ratio-based number assignments. The resultant ratio-scale data is appropriate for parametric statistical analysis. This method has been used successfully in a variety of usability activities at Oracle. It has proven to be efficient, sensitive, and highly effective for comparisons using usability as a differentiator.

INTRODUCTION

Oracle Corporation recently engaged in usability activities that found typical measures of usability ineffective for differentiation in traditional experimental hypothesis testing. An alternative was sought that could faithfully measure usability at the task level and was capable of discriminating under a variety of conditions. Using the psychophysical measurement technique of magnitude estimation to assess usability proved to be the metric that resolved our issues. Furthermore, an evaluation of the benefits of usability magnitude estimation shows potential to address many of the concerns and goals for measuring usability in general.

WHAT IS MAGNITUDE ESTIMATION?

Magnitude estimation is a psychophysical measurement method for assessing the *psycho*logical sensation of a *physical* stimulus. This is accomplished by having participants perform a number assignment procedure based on a subjective perception across a series of objects. The outcome from using magnitude estimation methodology is a ratio-scale continuum of the subjective perception under study. This ratio-scale continuum can then be used to make a variety of summary judgments about the objects under test, including parametric statistics.

Magnitude estimation is a highly documented method that is "extremely efficient", "one of the most frequently used psychological ratio scaling methods", "ideal for scaling large number of stimuli", and "superior to ordinal scale" (Gescheider, 1997).

Psychophysics has a long history in perceptual scaling of physical stimuli. Classic examples include the subjective assessment of line length with actual length, and the subjective assessment of brightness with actual amount of lumens emitted from a light source.

In addition to clear psychological and physical stimulus pairings, the method has proven flexible and robust enough to scale vague, multi-faceted perceptions with complex underlying physical stimuli. This is particularly compelling as multi-faceted perceptions that do not have a physical analog, particularly when produced from multidimensional stimuli, are very difficult to measure.

Gescheider (1997) cites successful uses of magnitude estimation in a variety of complex environments: trial evidence (physical stimulus) with guilt (perception); life events with emotional stress; and psychotic symptoms with severity of mental disorder. Furthermore, Snow (1998) and McGee (1998) established precedent for assessing perceptual phenomenon in visual interfaces by respectively measuring presence (the sense of 'being there') and cybersickness in helmet-mounted displays.

WHY USABILITY?

In software assessment the concept of usability is often used to judge the 'quality' of various software interface designs. Unfortunately, usability is a multi-faceted perceptual construct mediated by complex and varying physical stimuli. It is difficult to accurately and comprehensively assess usability so that meaningful comparisons can be made; i.e., "valid metrics that would be useful do not exist" (Lund, 1998).

Typical objective measures to assess usability, such as task completion rate, time, errors, subjective questionnaires, and other user comments are suitable to intuit areas of an interface needing improvement. In addition, they are frequently cited in summative evaluations looking to summarize and compare items of interest. However, Oracle recently initiated experiments to determine best designs in support of user interface standards and, as Lund (1998) suggests, found typical usability metrics insufficient for use as dependent measures seeking to statistically validate experimental hypotheses.

A key goal of usability evaluation is to assess items of interest across the full construct of usability and be able to discriminate among them accordingly. Thus, usability metrics

that *can* comprehensively assess usability and be used to differentiate items of interest are needed. The traditional measures employed for usability assessment do not meet these requirements (e.g., task completion rate, time, errors, and subjective questionnaires).

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 number of steps chosen for a given task easily manipulates time. Furthermore, task complexity can cause task completion rates to be so high and number of errors so low that meaningful discrimination cannot occur. Beyond confounds with tasks, performance metrics by nature assess narrow specific dimensions of usability that limit comprehensive usability assessment.

Subjective usability measures are typically large questionnaires that are limited to gross overall evaluations and infeasible 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, again leading to limited differentiability. Furthermore, subjective Likert rating scales are ordinal measures based on an assumed underlying continuum that generally have not been validated; i.e., a one-unit difference between values on a Likert scale has no meaning beyond the rank order of one value to the next.

Due to these limitations of traditional usability metrics, alternative usability metrics were sought that could comprehensively measure usability, be more resistant to task differences, and allow meaningful differentiation. Beyond Oracle's immediate need, Lund (1998) discusses how a reliable, valid usability metric could aid product comparisons, improve external validity of usability, validate user interface guidelines, allow more effective methodological research, and serve as the basis of a usability standard.

As described previously, magnitude estimation is a proven, sensitive, robust, and flexible measurement method for a variety of situations. Based on its characteristics and similar successful applications, magnitude estimation seemed to be a very credible solution for addressing the many concerns surrounding usability measurement.

HOW TO: ADMINISTRATION

The basic prerequisites for magnitude estimation are: 1) a series of targets for each participant to assess, and 2) an objective definition by which to estimate magnitude. Thus for usability, the first prerequisite is simply the tasks or trials within a usability evaluation. The second, which is the key to successful subjective magnitude estimation, is a definition of usability easily understood by participants.

Two corollary requirements to magnitude estimation measurement relate to minimum number of observations, both in targets to assess and participants making assessments. The minimum number of targets determines the breadth and reliability of the resulting usability scale. Two is an absolute minimum to make a comparison, while ten or more targets are recommended (Gescheider, 1997).

A sufficient number of participants are needed to statistically summarize the resulting data. There are many usability and experimental psychology philosophies on this topic. Recommendations in the magnitude estimation literature suggest ten participants (Gescheider, 1997). In practice, experiments have successfully employed magnitude estimation with as few as four participants (McGee, 1998); and many usability evaluations at Oracle have successfully used usability magnitude estimation with eight participants.

The main steps involved in actively administering usability magnitude estimation in a usability activity are:

1. Instruct the participant on the method.
2. Administer a standard practice magnitude estimation task.
3. Objectively define usability.
4. Objectively define the target (interface, task, trial, product, etc.).
5. Collect usability magnitude estimates after each target.

For Step 1, instructing the participant on the method, it is important to use clearly worded instructions for participants so that they quickly understand the method. In practice, it is a simple concept to understand, yet slightly different than typical rating assessments. The key tenets to reinforce with participants are to not use pre-defined scales (1-10, 0-100%, etc.) and to not impose anchors or limits on the range of ratio-based numbers used (except negative numbers which cannot be used in a ratio scale). The following instructions have been used in multiple tests at Oracle:

> "You will be asked to assign numbers to your subjective assessment of a series of 'targets'. These targets will be presented one at a time in random order. Your number assignments will be based on an objective definition provided to you (for example: brightness of lights, length of lines, or usability of software). When assessing the first target of a series, assign any arbitrary number that seems appropriate; however, be aware that this value will be the initial basis for subsequent comparisons. For all following targets, assign numbers in proportion to all previous assessments. For example, if your assessment of a target is twice as great as a previous target, assign a number twice as large. If your assessment of a target is one third of a previous target, assign a number one third as small. You may use any positive, non-zero numbers that seem appropriate (whole numbers, fractions, or decimals). There is no limit on how large or small the numbers can be that you assign to the targets. In other words, *do not use a pre-defined range of numbers at any time*; it is always possible for a target of greater or lower value to be shown than the ones you've seen previously."

For Step 2, administering a standard practice task, the goal is to have participants practice using the method, in addition to providing an analyzable post-test checkpoint to ensure magnitude estimation was used correctly. For the practice task, magnitude estimation is performed on a simple known

measurement standard (e.g., length of lines or size of circles – both have been used at Oracle). Participants are asked to use magnitude estimation to assign values for ten randomly ordered targets, presented one at a time.

For Step 3, objectively defining usability, the construct is formed from which a scale of usability will be made by participants. The goal is to rapidly and clearly inform participants about the concept of usability. The following definition was summarized from Oracle internal usability subscales and the ISO 9241 definition of usability:

> "Usability is your perception of how clear, easy to learn, easy to use, efficient, satisfying, etc. it is to accomplish a task with a particular system."

In practice, nearly all participants have pre-conceived notions about what usability means; however, the objective definition serves as a consistent baseline.

For Step 4, objectively defining the targets, the exact user interface units being tested need to be explained (interface, task, trial, product, etc.). Typically these are described in the general instructions of the study; however, the user interface target should be explicitly stated again after objectively defining usability within context of magnitude estimation.

For Step 5, collecting target estimates, participants are simply asked for their rating or estimate of usability after each target and their answer is recorded. This takes only a few seconds and is easily incorporated in between-task activities.

HOW TO: DATA REDUCTION

Assuming more than one participant is used, raw data collected by magnitude estimation is not ready for statistical analysis. Participants can construct scales of vastly different magnitude that need to undergo normalization before parametric analysis. The normalization procedure that magnitude estimation uses is geometric averaging. There are five basic steps to completing this procedure:

1. Collate the raw usability estimates per participant, per task (columns 1-3 of Table 1 -- Participant, Task, and U).
2. Calculate the log of each raw estimate and determine each participant's mean log score and the overall mean (column 4 -- Log U).
3. Determine each participant's offset from the overall mean by subtracting each participant's mean log score from the overall mean (column 5 -- Offset).
4. Add each participant's offset to each individual log score (column 6 -- Log U′).
5. Calculate the antilog of the normalized log scores of step 4 (column 7 -- U′).

The basic premise of geometric averaging is that log transformations are used to 'average' all participants' data onto one scale. The pertinent property of using log scores is that ratio information is preserved despite any addition or subtraction of a constant (e.g., the offset). The averaging procedure utilizes this property to construct a common scale conserving all participants' ratio information. The antilog step then reinstates the 'average' units of the original raw estimates, while preserving all original ratio information.

To verify, consider column 7 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 between tasks four and one (see column 3 -- U). The resulting ratio-scale continuum of U′ is now appropriate for parametric statistical analysis.

Table 1. Geometric averaging example. Note: For actual use, additional columns filled with participant and total Log U′ means allow for easier spreadsheet manipulation.

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
			$\overline{\chi}_1 = 1.35$			
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
			$\overline{\chi}_2 = 0.65$			
			$\overline{\chi}_{Total} = 1.00$			

U — Raw participant usability scores.
Offset — Overall Log U mean minus participant Log U mean.
Log U′ — Log U plus Offset.
U′ — Geometrically averaged usability scores (Antilog (Log U′)).

VALIDATION

Badia and Runyon (1982) state that if a measurement instrument assists in understanding and predicting behavior, then it is valid. They go on to discuss steps necessary to establish validity: repeated successful uses of a measure in a variety of settings, faithful comparisons with known accepted measures of the same construct, and predicted performance.

Thus far Oracle has completed ten usability activities, with additional investigations ongoing, that utilized usability magnitude estimation as the primary, or in some cases only usability metric. These activities have included formal usability benchmark tests, design alternative experimental hypothesis tests, diagnostic-oriented usability evaluations, and rapid iterative design tests. The interface domains have included desktop browsers, handheld devices (PDAs and cellular phones), and voice-interactive applications. The products tested have included 'front-end' applications such as email, calendar, and employee directory; and 'back-end' applications such as field service, inventory management, and human resources.

A variety of studies in the literature have shown that magnitude estimation measures, when defined appropriately to its construct, faithfully correlate with accepted measures where possible to assess (Meister, 1985; Gescheider, 1997; McGee, 1998). Usability magnitude estimation has shown the same property, notwithstanding the limitations of common usability metrics.

The first extensive use of usability magnitude estimation at Oracle examined correlations with typical performance usability metrics (McGee, Courage, and Nash, 2001). Usability magnitude estimation was significantly correlated with task completion time ($r = -.244$, $p < .001$), number of 'clicks' ($r = -.387$, $p < .000$), errors ($r = -.195$, $p < .011$), and assists ($-.193$, $p < .012$) (task completion rate was not analyzed as only a few tasks in the whole study were not completed). Correlation with simple Oracle internal subjective Likert scales showed relationships at $p < .10$. Other standardized subjective usability evaluations were not collected for this test.

Other than hypothesis testing that collected usability magnitude estimates and performance metrics simultaneously (i.e., the correlations cited above), explicit investigations predicting behavior based on prior measured usability magnitude estimation have not been conducted due to practical constraints of real usability testing.

DISCUSSION

Usability magnitude estimation has proven to be a success in a variety of usability activities at Oracle. Its use is rapidly increasing in frequency and importance. This is a direct result of the advantageous characteristics of usability magnitude estimation:

- High construct validity; i.e., actually measures underlying phenomenon of usability
- Efficient, fast, flexible administration
- Ratio level measurement scale data
- Precise measurement due to ease of administration and ratio-scale data
- Data appropriate for formative and summative statistical analysis
- Sensitivity to subtle differences between stimuli
- Ease of interpretation

While the tactical strengths of usability magnitude estimation are important, the ease of interpretation has implications beyond the usability work itself. Dumas and Redish (1999) cite the need to be able to communicate effectively with development teams and provide clear defensible usability statements. Usability magnitude estimation results can be instantiated in meaningful and defendable data-driven statements such as: *The usability of task A was 65% greater than task B...* and *The task sequence with the new feature had 25% higher usability than without.* Or if preferred, conclusions using the ratio information can be made: *Product A is rated 4 ½ times more usable than Product B.* In both cases, results are stated using very common measurement units that any executive, developer, or customer can easily understand (percentage and simple multiplication).

As an example, one development team within Oracle shrewdly requested shorter tasks in response to poor comparisons with other products on task completion time. However, when usability comparisons were highlighted by the usability magnitude estimation metric, they readily understood and accepted the issues.

FUTURE WORK

Oracle is pursuing a variety of extensions and uses of usability magnitude estimation in other usability activities. The most important extension is master scaling, a technique to standardize different sets of stimuli tested by different users, using the same objective definition (Gesheider, 1997). The resulting master usability scale appears to have the properties of the elusive universal measurement of usability alluded to by Lund (1998) (work in progress).

In addition to the master usability scale, the flexibility of usability magnitude estimation has allowed a quantitative component to be added to formative evaluation of very large design spaces that was not possible previously (work in progress).

Other work needed to further solidify the validity of the method includes predictive criterion testing, more rigorous validation of the objective definition of usability, validation with other usability metrics, and use of the method outside of Oracle.

CONCLUSIONS

Usability metrics are frequently labeled as 'objective' or 'subjective' with limited in-depth consideration of what makes a metric successful. Usability professionals need to have confidence that their conclusions are more science than snake oil when communicating results to developers, executives, and customers. Improved metrics for usability can increase the efficacy of our work and our understanding of the complex phenomenon of usability. Usability magnitude estimation is such a measure, one that has proven to be a highly effective, efficient, *and* satisfying for measuring usability.

REFERENCES

Dumas, J.S. & Redish, J.C. (1999). *A Practical Guide to Usability Testing, Revised Edition.* Intellect Books.

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

ISO 9241 Parts 1,2, 10-17 (1994-1996). Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs). International Standard.

Lund, A. (1998). The Need for A Standardized Set of Usability Metrics. *HFES Proceedings 1998, 42nd meeting,* pp. 688-691.

McGee, M. (1998). *Assessing Negative Side Effects in Virtual Environments*. Unpublished Masters thesis. Virginia Polytechnic Institute and State University. Blacksburg, VA.

McGee, M., Courage, C., & Nash, E. (2001). Mobile User Interface Design Tests: Three Experiments on Multiple Value Input on Internet-enabled Cell Phones. Internal Oracle Document.

Meister, D. (1985). *Behavioral Analysis and Measurement Methods*. John Wiley & Sons, Inc.

Pietro, B. & Runyon, R.P. (1982). *Fundamentals of Behavioral Research*. McGraw-Hill.

Snow, M. (1998). Empirical Models Based on Free-Modulus Magnitude Estimation of Perceived Presence in Virtual Environments. *Human Factors, Vol. 40*, No. 3, September 1998, pp. 386-402.