CHAPTER

Combined and Comparative Metrics

8

Usability data are building blocks. Each piece of usability data can be used to create new metrics. The raw usability data might be task completion rates, time-on-task, or self-reported ease of use. All of these usability data can be used to derive new metrics that were not previously available, such as an overall usability metric or a usability scorecard. Two ways to derive new usability metrics from existing data are (1) by combining more than one metric into a single usability measure and (2) by comparing existing usability data to expert or ideal results. We will review both methods in this chapter.

8.1 SINGLE USABILITY SCORES

In many usability tests, you collect more than one metric, such as task completion rate, task time, and perhaps a self-reported metric such as a System Usability Scale (SUS) score. In most cases, you don't care so much about the results for each of these metrics individually as you do about the total picture of the usability of the product as reflected by *all* of them. This section covers the various ways you can combine or represent different metrics to get an overall view of the usability of a product, or of different aspects of a product, perhaps as revealed by different tasks.

The most common question asked after a usability test is "How did it do?" The people who ask this question (often the product manager, developer, or other members of the project team) usually don't want to hear about task completion rates, task times, or questionnaire scores. They want an overall score of some type: Did it do well or not? How did it do in comparison to the last round of usability testing? Making these kinds of judgments in a meaningful way involves combining the metrics from a usability test into some type of single usability score. The challenge is figuring out how to combine scores from different scales appropriately (e.g., task completion rates in percentages and task times in minutes or seconds).

8.1.1 Combining Metrics Based on Target Goals

Perhaps the easiest way to combine different metrics is to compare each data point to a target goal and represent one single metric based on the percentage of participants who achieved a combined set of goals. For example, assume that the goal is for participants to successfully complete at least 80 percent of their tasks in no more than 70 seconds each on the average. Given that goal, consider the data in Table 8.1, which shows the task completion rate and average time per task for each of eight participants in a usability test. Also shown is an indication of whether each participant met the objective of completing at least 80 percent of the tasks in no more than 70 seconds.

Table 8.1 presents some interesting results. The average values for task completion (82 percent) and task time (67 seconds) would seem to indicate that the goals for this test were met. Even if you look at the number of participants who met the task completion goal (six participants, or 75 percent) or the task time goal (five participants, or 62 percent), you still find the results reasonably encouraging. However, the most appropriate way to look at the results is to see if each individual participant met the stated goal (i.e., the *combination* of completing at least 80 percent of the tasks in no more than 70 seconds). It turns out, as shown in the last column of the table, that only three, or 38 percent, of the participants actually met the goal. This demonstrates the importance of looking at individual participant data rather than just looking at averages.

This method of combining metrics based on target goals can be used with any set of metrics. The only real decision is what target goals to use. Target goals can be based on business goals and/or comparison to ideal performance. The math is easy (each person just gets a 1 or a 0), and the interpretation is easy to explain (the percentage of participants who had an experience that met the stated goal during the test).

Table 8.1 Sample Task Completion and Task Time Data

Participant Number	Task Completion	Task Time (sec)	Goal Met?
1	85%	68	1
2	70%	59	0
3	80%	79	0
4	75%	62	0
5	90%	72	0
6	80%	60	1
7	80%	56	1
8	95%	78	0
Average	**82%**	**67**	**38%**

8.1.2 Combining Metrics Based on Percentages

Although we are well aware that we should have measurable target goals for our usability tests, in practice we often don't have them. So what can we do to combine different metrics when we don't have target goals? One simple technique for combining scores on different scales is to convert each score to a percentage and then average them. For example, consider the data in Table 8.2, which show the results of a usability test with ten participants.

One way to get an overall sense of the results from this study is to first convert each of these metrics to a percentage. In the case of the number of tasks completed and the subjective rating, it's easy because we know the maximum ("best") possible value for each of those scores: There were 15 tasks, and the maximum possible subjective rating on the scale was 4. So we just divide the score obtained for each participant by the corresponding maximum to get the percentage.

In the case of the time data, it's a little trickier since there's no predefined "best" or "worst" time—the ends of the scale are not known beforehand. One way to handle this is to treat the fastest time obtained as the "best" (25 seconds) and then express the other times in relation to it. Specifically, you divide the shortest time by each time observed to convert it to a percentage. This way, the shortest time becomes 100 percent. If a given time was twice as long as the shortest, it

Table 8.2 Sample Data from a Usability Test

Participant Number	Time per Task (sec)	Tasks Completed (of 15)	Rating (0–4)
1	65	7	2.4
2	50	9	2.6
3	34	13	3.1
4	70	6	1.7
5	28	11	3.2
6	52	9	3.3
7	58	8	2.5
8	60	7	1.4
9	25	9	3.8
10	55	10	3.6

Note: Time per Task is the average time to complete each task, in seconds. Tasks Completed is the number of tasks (out of 15) that the participant successfully completed. Rating is the average of several 5-point subjective rating scales, where higher is better.

Table 8.3 Data from Table 8.2 Transformed to Percentages

Participant Number	Time	Tasks	Rating	Average
1	38%	47%	60%	48%
2	50%	60%	65%	58%
3	74%	87%	78%	79%
4	36%	40%	43%	39%
5	89%	73%	80%	81%
6	48%	60%	83%	64%
7	43%	53%	63%	53%
8	42%	47%	35%	41%
9	100%	60%	95%	85%
10	45%	67%	90%	67%

Note: For the Task Completion data, the score was divided by 15. For the Rating data, the score was divided by 4. For the Time data, the shortest time (25) was divided by the time obtained.

becomes 50 percent. Using that method of transforming the data, you get the percentages shown in Table 8.3.

Table 8.3 also shows the average of these percentages for each of the participants. If any one participant had successfully completed all the tasks in the shortest average time and had given the product a perfect score on the subjective rating scales, that person's average would have been 100 percent. On the other hand, if any one participant had failed to complete any of the tasks, had taken the longest time per task, and had given the product the lowest possible score on the subjective rating scales, that person's average would have been closer to 0 percent. (The average can't actually reach 0 percent because the time data expressed as a percentage can only approach 0 percent, not actually reach it.) Of course, rarely do you see either of those extremes. Like the sample data in Table 8.3, most participants fall between those two extremes. In this case, the averages range from a low of 39 percent (Participant 4) to a high of 85 percent (Participant 9), with an overall average of 62 percent.

So, if you had to give an "overall score" to the product whose test results are shown in Tables 8.2 and 8.3, you could say it scored 62 percent overall. Most people wouldn't be too happy with 62 percent. Many years of grades from school have probably conditioned most of us to think of a percentage that low as a "failing grade." But you should also consider how accurate that percentage is. Since it's an average based on the scores from ten different participants, you can construct a confidence interval for that average, as explained in section 2.4.4. The 95 percent

confidence interval in this case is 51 to 72 percent. Running more participants would probably give you a more accurate estimate of this value, whereas running fewer would probably have made it less accurate.

One thing to be aware of is that when we averaged the three percentages together (from the task completion data, task time data, and subjective ratings), we gave equal weight to each measure. In many cases, that is a perfectly reasonable thing to do, but sometimes the business goals of the product may indicate a different weighting. In this example, we're combining two performance measures (task completion and task time) with one self-reported measure (rating). By giving equal weight to each, we're actually giving twice as much weight to performance as to the self-reported measure. That can be adjusted by using weights in calculating the averages, as shown in Table 8.4. Each individual percentage is multiplied by its associated weight; these products are summed and that sum is divided by the sum of the weights (4 in this example).

In Table 8.4, the subjective rating is given a weight of 2, and each of the two performance measures is given a weight of 1. The net effect is that the subjective rating gets as much weight in the calculation of the average as the two performance measures together. The result is that these weighted averages for each participant tend to be closer to the subjective ratings than the equal-weight averages in Table 8.3. The exact weights you use for any given product should be determined by the business goals for the product. For example, if you're testing a website for use by the general public, and the users have many competitors' websites to choose from, you might want to give more weight to self-reported

Table 8.4 Calculation of Weighted Averages

Participant Number	Time	Weight	Tasks	Weight	Rating	Weight	Weighted Average
1	38%	1	47%	1	60%	2	51%
2	50%	1	60%	1	65%	2	60%
3	74%	1	87%	1	78%	2	79%
4	36%	1	40%	1	43%	2	40%
5	89%	1	73%	1	80%	2	81%
6	48%	1	60%	1	83%	2	68%
7	43%	1	53%	1	63%	2	55%
8	42%	1	47%	1	35%	2	40%
9	100%	1	60%	1	95%	2	88%
10	45%	1	67%	1	90%	2	73%

measures because you probably care more about the users' *perception* of the product than anything else.

On the other hand, if you're dealing with a safety-critical product such as an Automated External Defibrillator (AED), you probably want to give more weight to performance measures. You can use any weights that are appropriate for your situation, but remember to divide by the sum of those weights in calculating the weighted average.

To look at transforming one more set of metrics, consider the data shown in Table 8.5. In this case, the number of errors is listed, which would include specific ones, such as data-entry errors, the 12 participants made. Obviously, it is possible (and desirable) for a participant to make no errors, so the minimum possible is 0. But there's usually no predefined maximum number of errors that a participant could make. In a case like this, the best way to transform the data is to divide the number of errors obtained by the maximum number of errors and then subtract from 1. This is how the error percentages in Table 8.5 were obtained.

If any participant had no errors (optimum), her percentage would be 100 percent. The percentage for any with the highest number of errors would be 0 percent. Notice that in calculating any of these percentages, we always want higher percentages to be better—to reflect better usability. So in the case of errors, it makes more sense to think of the resulting percentage as an "accuracy" measure.

When transforming any usability metric to a percentage, the general rule is to first determine the minimum and maximum values that the metric can possibly have. In many cases this is easy; they are predefined by the conditions of the usability test. Here are the various cases you might encounter:

- If the minimum possible score is 0 and the maximum possible score is 100 (e.g., a SUS score), then you've already got a percentage.
- In many cases, the minimum is 0 and the maximum is known, such as the total number of tasks or the highest possible rating on a rating scale. In that case, simply divide the score by the maximum to get the percentage. (This is why it's generally easier to code rating scales starting with 0 as the worst value.)
- In some cases, the minimum is 0 but the maximum is not known, such as the example of errors. In that situation, the maximum would need to be defined by the data—the highest number of errors any participant made. Specifically, the number of errors would be transformed by dividing the number of errors obtained by the maximum number of errors any participant made and subtracting that from 1.
- Finally, in some cases neither the minimum nor maximum possible scores are predefined, as with time data. Assuming higher numbers are worse, as is the case with time data, you usually want to transform the data by dividing the lowest (best) score by the score obtained.

Table 8.5 Sample Data from a Usability Test

Participant Number	Tasks Completed (of 10)	Number of Errors	Satisfaction Rating (0–6)	Tasks	Accuracy	Satisfaction	Average
1	8	2	4.7	80%	60%	78%	73%
2	6	4	4.1	60%	20%	68%	49%
3	7	0	3.4	70%	100%	57%	76%
4	5	5	2.4	50%	0%	40%	30%
5	9	2	5.2	90%	60%	87%	79%
6	5	4	2.7	50%	20%	45%	38%
7	10	1	5.1	100%	80%	85%	88%
8	8	1	4.9	80%	80%	82%	81%
9	7	3	3.1	70%	40%	52%	54%
10	9	2	4.2	90%	60%	70%	73%
11	7	1	4.5	70%	80%	75%	75%
12	8	3	5.0	80%	40%	83%	68%

Note: Tasks Completed is the number of tasks (out of 10) that the participant successfully completed. Number of Errors is the number of specific errors that the participant made. Satisfaction Rating is on a scale of 0 to 6.

8.1.3 Combining Metrics Based on *z*-Scores

Another technique for transforming scores on different scales so that they can be combined is using *z*-scores. (See, for example, Martin & Bateson, 1993, p. 124.) These are based on the normal distribution and indicate how many units any given value is above or below the mean of the distribution. When you transform a set of scores to their corresponding *z*-scores, the resulting distribution by definition has a mean of 0 and standard deviation of 1. This is the formula for transforming any raw score into its corresponding *z*-score:

$$z = (x - \mu)/\sigma$$

where

x = the score to be transformed
μ = the mean of the distribution of those scores
σ = the standard deviation of the distribution of those scores

This transformation can also be done using the STANDARDIZE function in Excel.

The data in Table 8.2 could also be transformed using *z*-scores, as shown in Table 8.6 (on page 200). For each original score, the *z*-score was determined by subtracting the mean of the score's distribution from the original score and then dividing by the standard deviation. This *z*-score tells you how many standard deviations above or below the mean that score is.

Table 8.6 shows the mean and standard deviation for each set of *z*-scores, which should always be 0 and 1, respectively. Note that in using *z*-scores, we didn't have to make any assumptions about the maximum or minimum values that any of the scores could have. In essence, we let each set of scores define its own distribution and rescaled them so those distributions would each have a mean of 0 and a standard deviation of 1.

In this way, when they are averaged together, each of the *z*-scores makes an equal contribution to the average *z*-score. Notice that when averaging the *z*-scores together, each of the scales must be going in the same direction—in other words, higher values should always be better. In the case of the time data, the opposite is almost always true. Since *z*-scores have a mean of 0, this is easy to correct simply by multiplying the *z*-score by (−1) to reverse its scale.

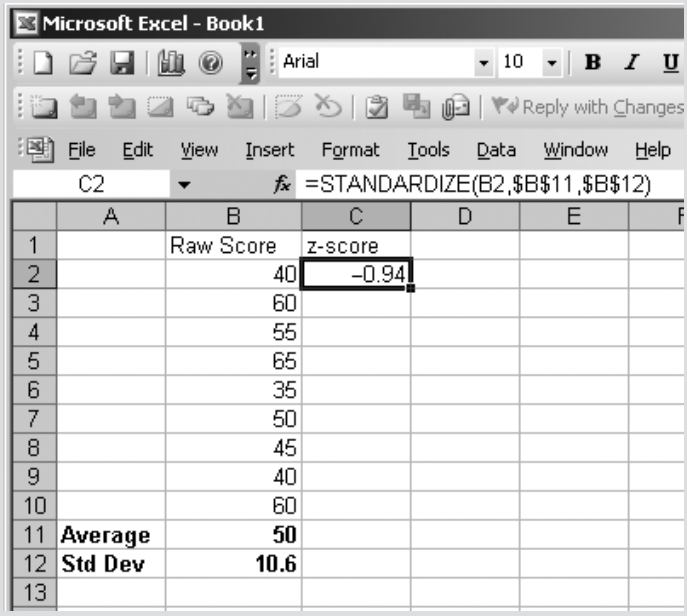
If you compare the *z*-score averages in Table 8.6 to the percentage averages in Table 8.3, you will find that the ordering of the participants based on those averages is nearly the same: Both techniques yield the same top three participants (9, 5, and 3) and the same bottom three participants (4, 8, and 1).

One disadvantage of using *z*-scores is that you can't think of the overall average of the *z*-scores as some type of overall usability score, since by definition that

STEP-BY-STEP GUIDE TO CALCULATING *z*-SCORES

Here are the steps for transforming any set of raw scores (times, percentages, clicks, whatever) into *z*-scores:

1. Enter the raw scores into a single column in Excel.
2. Calculate the average and standard deviation for this set of raw scores using "=AVERAGE(range of raw scores)" and " =STDEV(range of raw scores)."
3. In the cell to the right of the first raw score, enter the formula "=STANDARDIZE(<raw score>,<average>,<std dev>)," where <raw score> is replaced by a reference to the first raw score to be converted to a *z*-score, <average> is a reference to the average of the raw scores, and <std dev>is a reference to the standard deviation of the raw scores. The figure that follows shows an example. Note that the references to the cells containing the average (B11) and standard deviation (B12) include "$" in them. That's so that the references are "locked" and don't change as we copy the formula down the rows.
4. Copy this "standardize" formula down as many rows as there are raw scores.
5. As a double-check, copy the formulas for the average and standard deviation over to the *z*-score column. The average should be 0, and the standard deviation should be 1.



	A	B	C	D	E
1		Raw Score	z-score		
2		40	-0.94		
3		60			
4		55			
5		65			
6		35			
7		50			
8		45			
9		40			
10		60			
11	**Average**	**50**			
12	**Std Dev**	**10.6**			
13					

Table 8.6 Sample Data from Table 8.2 Transformed Using *z*-Scores

Participant Number	Time per Task (sec)	Tasks Completed (of 15)	Rating (0–4)	*z*-Time	*z*-Time (–1)	*z*-Tasks	*z*-Rating	Average
1	65	7	2.4	0.98	–0.98	–0.91	–0.46	–0.78
2	50	9	2.6	0.02	–0.02	0.05	–0.20	–0.06
3	34	13	3.1	–1.01	1.01	1.97	0.43	1.14
4	70	6	1.7	1.30	–1.30	–1.39	–1.35	–1.35
5	28	11	3.2	–1.39	1.39	1.01	0.56	0.99
6	52	9	3.3	0.15	–0.15	0.05	0.69	0.20
7	58	8	2.5	0.53	–0.53	–0.43	–0.33	–0.43
8	60	7	1.4	0.66	–0.66	–0.91	–1.73	–1.10
9	25	9	3.8	–1.59	1.59	0.05	1.32	0.98
10	55	10	3.6	0.34	–0.34	0.53	1.07	0.42
Mean	**49.7**	**8.9**	**2.8**	**0.0**	**0.0**	**0.0**	**0.00**	**0.00**
Standard Deviation	**5.6**	**2.1**	**0.8**	**1.0**	**1.0**	**1.0**	**1.00**	**0.90**

overall average will be 0. So when would you want to use *z*-scores? They are useful mainly when you want to compare one set of data to another, such as data from iterative usability tests of different versions of a product, data from different groups of participants in the same usability session, or data from different conditions or designs within the same usability test.

For example, consider the data shown in Figure 8.1 from Chadwick-Dias, McNulty, and Tullis (2003), which shows *z*-scores of performance for two iterations of a prototype. This research was studying the effects of age on performance in using a website. Study 1 was a baseline study. Based on their observations of the participants in Study 1, and especially the problems encountered by the older participants, the authors of the study made changes to the prototype and then conducted Study 2 with a new group of participants. The *z*-scores were equal-weighted combinations of task time and task completion rate.

It's important to understand that the *z*-score transformations were done using the *full set* of data from Study 1 and Study 2 combined. They were then plotted appropriately to indicate from which study each *z*-score was derived. The key finding was that the performance *z*-scores for Study 2 were significantly higher than the performance *z*-scores for Study 1, and the effect was the same regardless of age (as reflected by the fact that the two lines are parallel to each other). If the

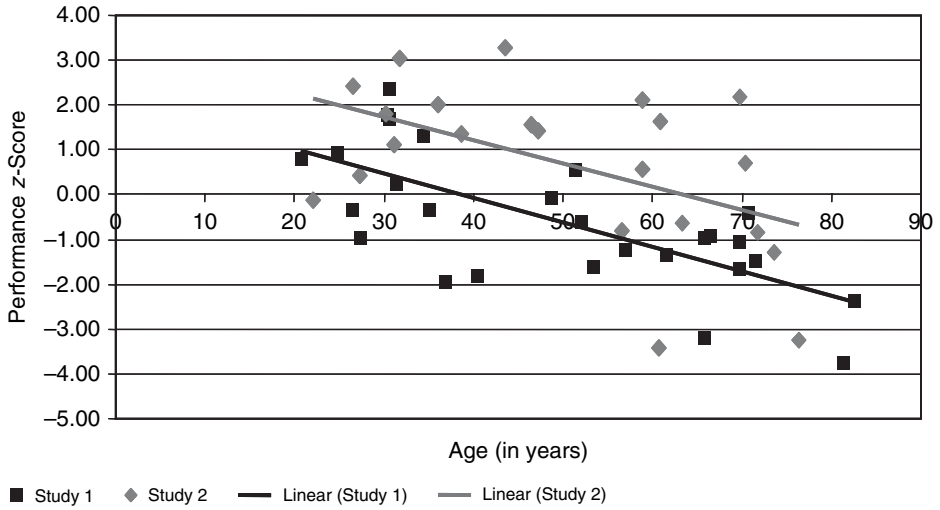


FIGURE 8.1

Data showing performance *z*-scores from two studies of a prototype with participants over a wide range of ages. The performance *z*-score was an equal-weighted combination of task time and task completion rate. Changes were made to the prototype between Study 1 and Study 2. The performance *z*-scores were significantly better in Study 2, regardless of the participant's age. *Source:* Adapted from Chadwick-Dias et al. (2003); used with permission.

z-score transformations had been done *separately* for Study 1 and Study 2, the results would have been meaningless because the means for Study 1 and Study 2 would both have been forced to 0 by the transformations.

8.1.4 Using SUM: Single Usability Metric

Jeff Sauro and Erika Kindlund (2005) have developed a quantitative model for combining usability metrics into a single usability score. Their focus is on task completion, task time, error counts per task, and post-task satisfaction rating (similar to ASQ described in section 6.3.2). Note that all of their analyses are at the task level, whereas the previous sections have described analyses at the "usability test" level. At the task level, task completion is typically a binary variable for each participant: that person either completed the task successfully or did not. At the usability-test level, task completion, as we have seen in the previous sections, indicates how many tasks each person completed, and it can be expressed as a percentage for each participant.

Sauro and Kindlund used techniques derived from Six Sigma methodology (e.g., Breyfogle, 1999) to standardize their four usability metrics (task completion, time, errors, and task rating) into a Single Usability Metric (SUM). Conceptually, their techniques are not that different from the *z*-score and percentage transformations described in the previous sections. In addition, they used Principal Components Analysis to determine if all four of their metrics were significantly contributing to the overall calculation of the single metric. They found that all four were significant and, in fact, that each contributed about equally. Consequently, they decided that each of the four metrics (once standardized) should contribute equally to the calculation of the SUM score.

An Excel spreadsheet for entering the data from a usability test and calculating the SUM score is available from Jeff Sauro's "Measuring Usability" website at *http://www.measuringusability.com/SUM/*. For each task and each participant in the usability test, you must enter the following:

- Whether the participant completed the task successfully (0 or 1).
- Number of errors committed on that task by that participant. (You also specify the number of error opportunities for each task.)
- Task time in seconds for that participant.
- Post-task satisfaction rating, which is an average of three post-task ratings on 5-point scales of task ease, satisfaction, and perceived time—similar to ASQ.

After entering these data for all the tasks, the spreadsheet standardizes the scores and calculates the overall SUM score and a confidence interval for each task. Standardized data for each task for ten participants and six tasks is illustrated in Table 8.7. Notice that a SUM score is calculated for each task, which allows for overall comparisons of tasks. In these sample data, the

Table 8.7 Sample Standardized Data from a Usability Test

Task	SUM Low	SUM Mean	SUM High	Completion	Satisfaction	Time	Errors
Reserve a room	62%	75%	97%	81%	74%	68%	76%
Find a hotel	38%	58%	81%	66%	45%	63%	59%
Check room rates	49%	66%	89%	74%	53%	63%	74%
Cancel reservation	89%	91%	99%	86%	91%	95%	92%
Check restaurant hours	22%	46%	68%	58%	45%	39%	43%
Get directions	56%	70%	93%	81%	62%	66%	71%
Overall SUM	**53%**	**68%**	**88%**				

participants did best on the "Cancel reservation" task and worst on the "Check restaurant hours" task. An overall SUM score, 68 percent in this example, is also calculated, as is a 90 percent confidence interval (53 to 88 percent), which is the average of the confidence intervals of the SUM score for each task.

8.2 USABILITY SCORECARDS

An alternative to combining different metrics to derive an overall usability score is to graphically present the results of the metrics in a summary chart. This type of chart is often called a Usability Scorecard. The goal is to present the data from the usability test in such a way that overall trends and important aspects of the data can be easily detected, such as tasks that were particularly problematic for the participants. If you have only two metrics that you're trying to represent, a simple combination graph from Excel may be appropriate. For example, Figure 8.2 shows the task completion rate and subjective rating for each of ten tasks in a usability test.

The combination chart in Figure 8.2 has some interesting features. It clarifies which tasks were the most problematic for participants (Tasks 4 and 8) because they have the lowest values on both scales. It's obvious where there were significant disparities between the task completion data and the subjective ratings, such as Tasks 9 and 10, which had only moderate task completion rates but the highest subjective ratings. Finally, it's easy to distinguish the tasks that had reasonably high values for both metrics, such as Tasks 3, 5, and 6.

This type of combination chart works well if you have only two metrics to represent, but what if you have more? One way of representing summary data for

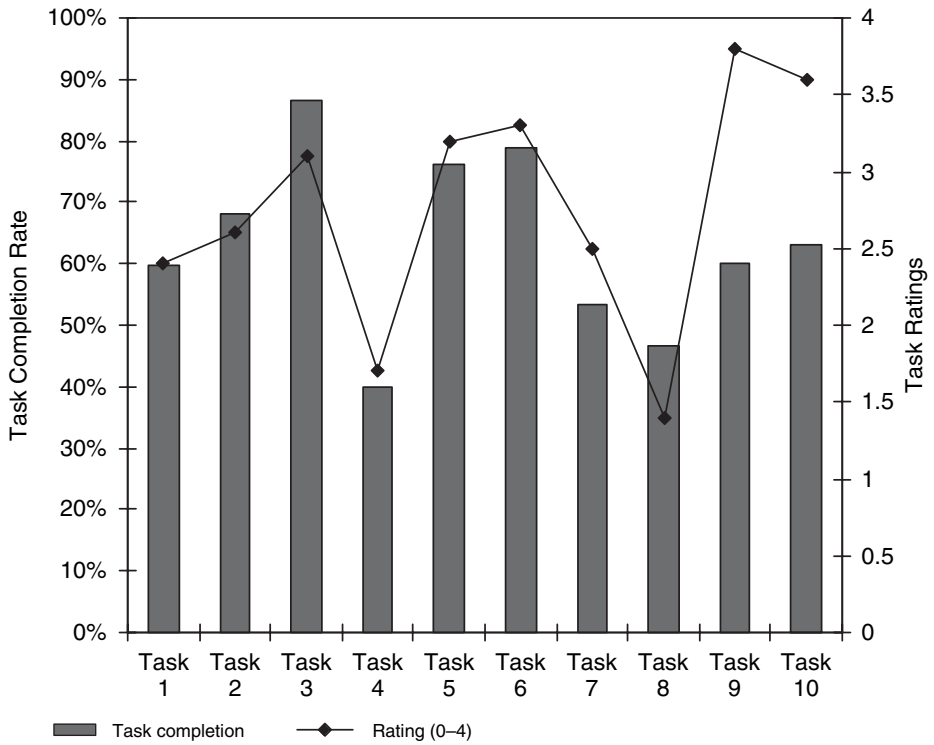


FIGURE 8.2

Sample combination column and line chart for ten tasks. Task completion data is shown via the columns and labeled on the *left* axis. Subjective rating is shown via the lines and labeled on the *right* axis.

three or more metrics is with radar charts (which were also illustrated in Chapter 6). Figure 8.3 shows an example of a radar chart for summarizing the results of a usability test with five factors: task completion, page visits, accuracy (lack of errors), satisfaction rating, and usefulness rating. In this example, although task completion, accuracy, and usefulness were relatively high (good), page visits and satisfaction were relatively low (poor).

Although radar charts can be useful for a high-level view, it's not really possible to represent task-level information in them. The example in Figure 8.3 averaged the data across the tasks. What if you want to represent summary data for three or more metrics but also maintain task-level information? One technique for doing that is using what are called Harvey Balls. A variation on this technique has been popularized by *Consumer Reports*. For example, consider the data shown earlier in Table 8.7, which presents the results for six tasks in a usability test, including

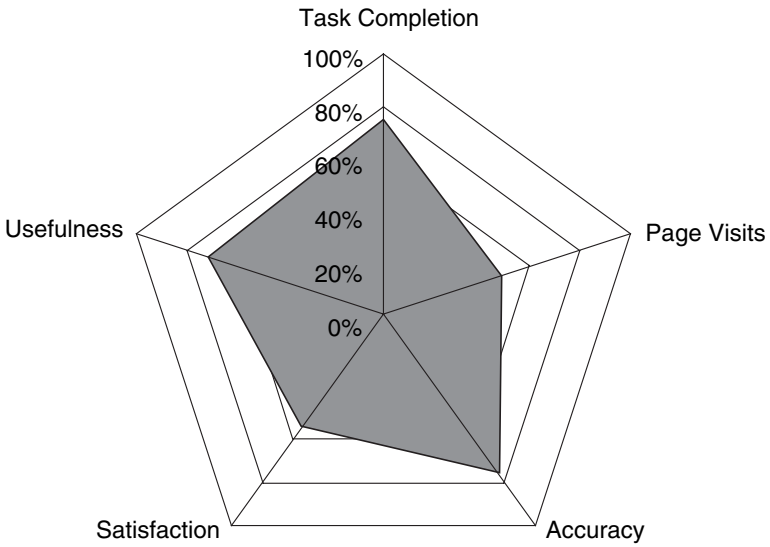


FIGURE 8.3

Sample radar chart summarizing task completion, page visits, accuracy (lack of errors), satisfaction rating, and usefulness rating from a usability test. Each score has been transformed to a percentage using the techniques outlined earlier in this chapter.

task completion, time, satisfaction, and errors. These data are summarized in the comparison chart shown in Figure 8.4. This type of comparison chart allows you to see at a glance how the participants did for each of the tasks (by focusing on the rows) or how the participants did for each of the metrics (by focusing on the columns).

WHAT ARE HARVEY BALLS?

Harvey Balls are small, round pictograms typically used in a comparison table to represent values for different items:

They're named for Harvey Poppel, a Booz Allen Hamilton consultant who created them in the 1970s as a way of summarizing long tables of numeric data. There are five levels, progressing from an open circle to a completely filled circle. Typically, the open circle represents the worst values, and the completely filled circle represents the best values. Links to images of Harvey Balls of different sizes can be found on our website, *www.MeasuringUserExperience.com*. Harvey Balls shouldn't be confused with Harvey Ball, who was the creator of the smiley face ☺!

Task	SUM Score	Completion	Satisfaction	Time	Errors
Cancel reservation	91%				
Reserve a room	75%				
Get directions	70%				
Check room rates	66%				
Find a hotel	58%				
Check restaurant hours	46%				

FIGURE 8.4

Sample comparison chart using the data from Table 8.7. The tasks have been ordered by their SUM score, starting with the highest. For each of the four standardized scores (task completion, satisfaction, task time, and errors), the value has been represented by a coded circle (known as a Harvey Ball), as shown in the key.

8.3 COMPARISON TO GOALS AND EXPERT PERFORMANCE

Although the previous section focused on ways to summarize usability data without reference to an external standard, in some cases you may have one that can be used for comparison. The two main flavors of an external standard are predefined goals and expert, or optimum, performance.

8.3.1 Comparison to Goals

Perhaps the best way to assess the results of a usability test is to compare those results to goals that were established before the test. These goals may be set at the task level or at an overall level. Goals can be set for any of the metrics we've discussed, including task completion, task time, errors, and self-reported measures. Here are some examples of task-specific goals:

- At least 90 percent of representative users will be able to successfully reserve a suitable hotel room.
- At least 85 percent of representative users will be able to open a new account online within ten minutes.

- At least 95 percent of new users will be able to purchase their chosen product online within five minutes of selecting it.

Similarly, examples of overall goals could include the following:

- Users will be able to successfully complete at least 90 percent of their tasks.
- Users will be able to complete their tasks in less than three minutes each, on the average.
- Users will give the application an average SUS rating of at least 80 percent.

Typically, usability goals address task completion, time, accuracy, and/or satisfaction. The key is that the goals must be measurable. You need to be able to determine whether the data in a given situation support the attainment of the goal. For example, consider the data in Table 8.8.

The data in the table show how many of the ten participants in the usability test actually completed each of the tasks and reached the stated goal. In this case, the goal is the same for each task (nine of ten participants), but the goals could be different for each task. One way of representing the data could be by showing the comparison of the actual number who completed each task as a percentage of the goal, as shown in Figure 8.5. This makes it easier to spot the tasks where participants had the most trouble (i.e., Tasks 5 and 10). Of course, this technique could be used to represent the percentage of participants who met any particular objective (e.g., time, errors, SUS rating, etc.) at either the task level or the overall level.

Table 8.8 Sample Task Completion Data and Goals

Task	Actual Number of Participants Who Completed	Goal
1	9	9
2	8	9
3	9	9
4	7	9
5	5	9
6	8	9
7	9	9
8	8	9
9	7	9
10	6	9

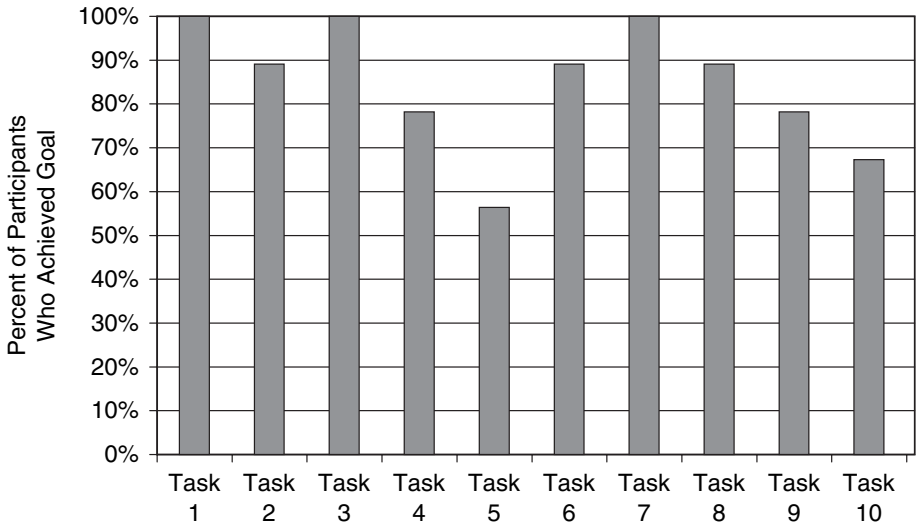


FIGURE 8.5

Alternative representation of the data in Table 8.8, showing the percentage of goal attainment for each task.

8.3.2 Comparison to Expert Performance

An alternative to comparing the results of a usability test to predefined goals is to compare the results to the performance of an "expert." The best way to determine the expert performance level is to have one or more presumed "experts" actually perform the tasks and to measure the same things that you're measuring in the usability test. Obviously your "experts" really need to be experts—people with subject-matter expertise, in-depth familiarity with the tasks, and in-depth familiarity with the product, application, or website being tested. And your data will be better if you can average the performance results from more than one expert.

Comparing the results of a usability test to the results for experts allows you to compensate for the fact that certain tasks may be inherently more difficult or take longer, even for an expert. The goal, of course, is to see how close the performance of the participants in the test actually comes to the performance of the experts.

Although you could theoretically do a comparison to expert performance for any performance metric, it's probably most common to do so for time data. With task completion data, the usual assumption is that a true expert would be able to perform all the tasks successfully. Similarly, with error data the assumption is that an expert would not make any errors. But even an expert would require some amount of time to perform the tasks. For example, consider the task time data in Table 8.9; it shows the average actual time per task, the expert time per task, and the ratio of expert to actual time.

Table 8.9 Sample Time Data from Ten Tasks

Task	Actual Time	Expert Time	Expert/Actual
1	124	85	69%
2	101	50	50%
3	89	70	79%
4	184	97	53%
5	64	40	63%
6	215	140	65%
7	70	47	67%
8	143	92	64%
9	108	98	91%
10	92	60	65%
Averages	**119**	**78**	**66%**

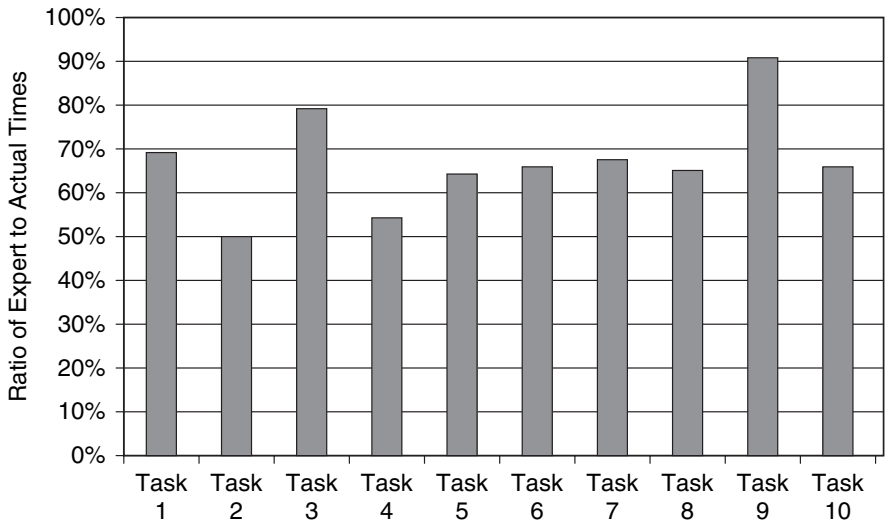


FIGURE 8.6

Graph of the ratio of the expert to actual times from Table 8.9.

Graphing the ratio of expert to actual times, as shown in Figure 8.6, makes it easy to spot the tasks where the test participants did well in comparison to the experts (Tasks 3 and 9) and the tasks where they did not do so well (Tasks 2

and 4). Notice that the average ratio of expert to actual performance—66 percent in this example—can also be used as an overall "usability score" for a usability test.

8.4 SUMMARY

Some of the key takeaways from this chapter are as follows.

1. An easy way to combine different usability metrics is to determine the percentage of participants who achieve a combination of goals. This tells you the overall percentage of participants who had a good experience with your product (based on the target goals). This method can be used with any set of metrics and is easily understood by management.
2. One way of combining different metrics into an overall usability score is to convert each of the metrics to a percentage and then average them together. This requires being able to specify, for each metric, an appropriate minimum and maximum score.
3. Another way to combine different metrics is to convert each metric to a *z*-score and then average them together. Using *z*-scores, each metric gets equal weight when they are combined. But the overall average of the *z*-scores will always be 0. The key is in comparing different subsets of the data to each other, such as data from different iterations, different groups, or different conditions.
4. The SUM technique is another method for combining different metrics, specifically task completion, task time, errors, and task-level satisfaction rating. The method requires entry of individual task and participant data for the four metrics. The calculations yield a SUM score, as a percentage, for each task and across all tasks, including confidence intervals.
5. Various types of graphs and charts can be useful for summarizing the results of a usability test in a usability scorecard. A combination line and column chart is useful for summarizing the results of two metrics for the tasks in a test. Radar charts are useful for summarizing the results of three or more metrics overall. A comparison chart using Harvey Balls to represent different levels of the metrics can effectively summarize the results for three or more metrics at the task level.
6. Perhaps the best way to determine the success of a usability test is to compare the results to a set of predefined usability goals. Typically these goals address task completion, time, accuracy, and satisfaction. The percentage of participants whose data met the stated goals can be a very effective summary.
7. A reasonable alternative to comparing results to predefined goals, especially for time data, is to compare the actual performance results to the results for experts. The closer the actual performance is to expert performance, the better.