COMBINE and COMPARATIVE METRICS

Tullis, T., & Albert, B. (2013). *Measuring the User Experience 2nd Edition.* Waltham, MA: Elsevier.

SINGLE USABILITY SCORES

Mendapatkan hasil pengujian performa produk secara keseluruhan, kombinasi dari beberapa pengukuran

COMBINING METRICS : TARGET GOALS

- Kombinasi ukuran untuk melihat ketercapaian sasaran
- Contoh : kombinasi ukuran task completion dan ukuran task time untuk melihat perfoma partisipan dalam gunakan produk. Apakah partisipan mencapai sasaran yang ditetapkan? Jika ya → 1, jika tidak → 0

Target Goal = TC min 80%, TT < 70 sec

Participant #	Task Completion	Task Time (secs)	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%

Table 8.1 Sample task completion and task time data from eight participants[a].
[a]Also shown are averages for task completion and time and an indication of whether each participant met the objective of completing at least 80% of the tasks in no more than 70 seconds.

HASIL PENGUJIAN

- Rata-rata Task Completion 82% - lebih dari tujuan
- Rata-rata Task time 67 detik – dianggap baik karena kurang dari batas (70 detik)
- Keduanya tidak perlihatkan keberhasilan tiap partisipan mencapai tujuan : minimal 80% tugas selesai dalam waktu kurang dari 70 detik
- Goal met /ketercapaian tujuan perlihatkan persentase partisipan yang sukses 38%

COMBINING METRICS : PERCENTAGES

- Saat tidak ada TARGET GOAL : Tiap skala/ukuran diubah dalam persentase, lalu hitung rata-rata
- Ukuran jumlah completed task dan subjective rating nilai maksimumnya adalah jumlah total tugas dan rating tertinggi
- Ukuran time : nilai terbaiknya adalah waktu tercepat dan nilai lain dibandingkan dengan waktu tercepat.

Data Sampel Tes Usability 10 Partisipan

Participant #	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

Table 8.2 Sample data from a usability test with 10 participants*.
*Time per task is the average time to complete each task, in seconds. Tasks completed are number of tasks (out of 15) that the user completed successfully. Rating is the average of a five-point task ease rating for each task, where higher is better.

Contoh Perhitungan

- Partisipan #1
  - Time per task (sec) : (70-65)/(70-25)x100% = 11 %
  - Task Completed (0f 15) : 7/15 X 100% =47%
  - Rating : 2,4/4 x 100% = 48%
- Dimana ..
  - Time per task (sec) : 70 waktu terlama, 25 waktu tercepat dan 65 waktu yang diperoleh
  - Task completed : 15 adalah jumlah total tugas
  - Rating : 4 adalah nilai maksimum rating

Tabel Transformasi Data ke Persentase

Participant #	Time	Tasks	Rating	Average
1	11%	47%	60%	39%
2	44%	60%	65%	56%
3	80%	87%	78%	81%
4	0%	40%	43%	28%
5	93%	73%	80%	82%
6	40%	60%	83%	61%
7	27%	53%	63%	48%
8	22%	47%	35%	35%
9	100%	60%	95%	85%
10	33%	67%	90%	63%

Table 8.3 Data from Table 8.2 transformed to percentages[a].
[a]For task completion data, the score was divided by 15. For rating data, the score was divided by 4. For time data, the difference between the longest time (70) and the observed time was divided by the difference between longest (70) and shortest (25) times.

Hasil

- Terendah 28%, tertinggi 85%
- Rata-rata keseluruhan 57.8% atau 58%
- Nilai 58% tidak mewakili 10 partisipan, karena itu gunakan 90% convidence interval : ± 11% sehingga convidence interval dari 47% sampai 69%
- 47% ------(-11)----- 58% -------(+11)-----69%

Dengan Bobot untuk Ukuran

Participant #	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%

Table 8.4 Calculation of weighted averages[a].
[a]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).

Perhitungan dengan Bobot

- Ukuran diberi bobot lebih karena dianggap penting dan menjadi pembeda
- Contoh :
  - Bobot time dan task (performa) > rating → sasaran bisnis : kecepatan dan akurasi penggunaan produk
  - Bobot Rating (berasal dari Self-reported metrics) > performa → sasaran bisnis adalah persepsi pengguna
- Bobot diselaraskan dengan business goal (sasaran bisnis) untuk produk
- Hasil cenderung mendekati bobot terbesar

COMBINING METRICS : PERCENTAGES (2)

Participant #	Tasks Completed (of 10)	# of Page Visits (min = 20)	Satis-faction Rating (0–6)	Useful-ness Rating (0–6)	Tasks	Page Visits	Satis-faction	Useful-ness	Average
1	8	32	4.7	3.9	80%	63%	78%	65%	71%
2	6	41	4.1	3.8	60%	49%	68%	63%	60%
3	7	51	3.4	3.7	70%	39%	57%	62%	57%
4	5	62	2.4	2.3	50%	32%	40%	38%	40%
5	9	31	5.2	4.2	90%	65%	87%	70%	78%
6	5	59	2.7	2.9	50%	34%	45%	48%	44%
7	10	24	5.1	4.8	100%	83%	85%	80%	87%
8	8	37	4.9	4.3	80%	54%	82%	72%	72%
9	7	65	3.1	2.5	70%	31%	52%	42%	49%

Table 8.5 Sample data from a usability test with nine participants[a].
[a]Tasks completed are the number of tasks (out of 10) that the user completed successfully. Number of page visits is the total number of web pages that the user visited in attempting the tasks.(Typically, each revisit to the same page is counted as another visit.) The two ratings are average subjective ratings of satisfaction and usefulness, each on a seven-point scale (0–6)

Perhitungan

- Task completed : total 10 tugas → nilai max 10 = 100%
- # Page visits : jumlah halaman yang dikunjungi untuk selesaikan tugas. Jumlah minimal : 20 (nilai terbaik) → jumlah_halaman/20*100%
- Subjective rating : satisfaction dan usefullness dari self-reported metrics, max : 6 → rating/6*100%
- Average = jumlah persentase ukuran /4

COMBINING METRICS: PERCENTAGES (3)

Participant #	Tasks Completed (of 10)	# of Errors	Satis-faction Rating (0–6)	Tasks	Accuracy	Satis-faction	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%

Table 8.6 Sample data from a usability test with 12 participants[a].
[a]Tasks completed are the number of tasks (out of 10) that the user completed successfully. Number of errors is the number of specific errors that the user made, such as data-entry errors. Satisfaction rating is on a scale of 0 to 6.

Perhitungan

- Task completed : total 10 tugas → nilai max 10 = 100%
- #Error: jumlah error yang dilakukan, min =0 →(1-jml_error/jml_error_max)*100%
- Jml_error_max diperoleh dari data
- Subjective rating : satisfaction self-reported metrics, max : 6 → rating/6*100%
- Average = jumlah persentase ukuran /3

ATURAN

- Tentukan nilai min –max yang mungkin terjadi
- Jika min=0, max=100 → 0% - 100%
- Jika min= 0, max=nilai_max → data/nilai_max
- Jika min=0, max=? → nilai max = nilai data tertinggi
- Jika min=?, max=? → (data_max-data)/(data_max-data_min)*100%

Data Outlier

- Jika ada data yang mencolok perbedaannya, misalnya dari Tabel 8.6, error partisipan 4 = 20
- Gunakan nilai mean dan standar deviasi dari data dan nilai data yang lebih dari 2 atau 3 kali standar deviasi
- Jika nilai data di luar 2 atau 3 kali std deviasi dari mean, maka nilai itu dianggap outlier, lalu tidak dianggap

COMBINING METRICS : z-Scores

- Ubah nilai ke persentase gunakan z-score atau fungsi STANDARDIZE pada excel
- =STANDARDIZE(A1, AVERAGE(A:A), STDEV(A:A)) dengan data di kolom A.
- Tidak ada asumsi nilai Min dan Max dari data

Tabel Data – Z Scores

Participant #	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				0.0	0.0	0.0	0.00	0.00
Standard deviation				1.0	1.0	1.0	1.00	0.90

Table 8.7 Sample data from Table 8.2 transformed using *z* scores*.
*For each original score, the *z* score was determined by subtracting the mean of the score's distribution from it and then dividing by the standard deviation. This *z* score tells you how many standard deviations above or below the mean that score is. Since you need all the scales to have higher numbers better, the scale of the *z* scores of times is reversed by multiplying by (–1).

Aturan

- Z score haruskan nilai ke arah yang sama, misal NILAI TERTINGGI ADALAH TERBAIK
- Untuk ubah arah kalikan data dengan -1
- Hasil z scores dan percentage dengan asumsi hampir sama.

Aturan

- Z scores digunakan untuk bandingkan satu data set ke data set lain. Misalnya:
  - data usability test dari produk yang beda versi
  - Data usability test yang sama dari beberapa grup pengguna
  - Data usability test yang sama dari beberapa kondisi atau desain yang berbeda
- Sampel yang layak, misalnya 10 orang per kondisi

Perbandingan – Z Scores

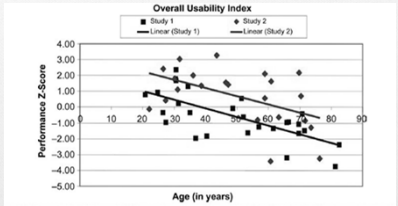


Figure 8.1 Data showing performance *z* scores from two studies of a prototype with participants over a wide range of ages. The performance *z* score was an equal-weighted combination of task time and task completion rate. Changes were made to the prototype between Study 1 and Study 2. The performance *z* scores were significantly better in Study 2, regardless of the participant's age. Adapted from Chadwick-Dias et al. (2003); used with permission.

USING SUM: Single Usability Metrics

- Dikembangkan oleh Jeff Sauro dan Erika Kindlund (2005)
- Fokus pengukuran SUM : TASK COMPLETION, TASK TIME, ERROR, SATISFACTION RATING
- Teknik tidak beda dari z-score dan percentage transformation, dengan tambahan Principal Components Analysis

- PCA : korelasi antar variabel untuk melihat kontribusi 4 variabel pada perhitungan single metric
- Online tool : www.usabilityscorecard.com
- SUM score dihitung untuk tiap tugas untuk perbandingan tiap semua tugas.

Tabel Data – SUM

	SUM						
Task	Low	Mean	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%				

Table 8.8 Sample standardized data from a usability test[a].
[a]After entering data for each participant and each task, these are the standardized scores calculated by SUM, including an overall SUM score and a confidence interval for it.

Graph Tool - SUM

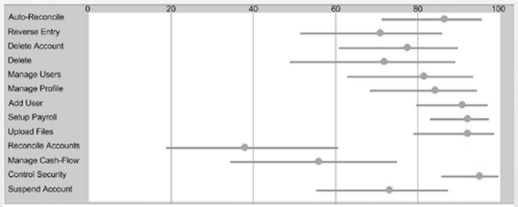


Figure 8.2 Sample graph of SUM scores from http://www.usabilityscorecard.com/. The tasks of this usability test are listed down the left. For each task, the orange circle shows the mean SUM score and bars show the 90% confidence interval for each. In this example, it's apparent that the "Reconcile Accounts" and "Manage Cash-Flow" tasks are the most problematic.

USABILITY SCORECARDS

- Penyajian hasil pengukuran lewat Summary Chart
- Sasaran : sajikan data dari usability test sehingga trend dan aspek penting secara keseluruhan terdeteksi
- Contoh :
  - Tugas yang bermasalah untuk pengguna
  - Dua ukuran : task completion dan task ease rating

SUMMARY CHART

Figure 8.3 A sample combination column and line chart for 10 tasks. Task rating is shown via the columns and labeled on the right axis. Task success is shown via the lines and is labeled on the left axis.

Penjelasan SUMMARY CHART

- Task 4 dan 8 : tugas yang paling bermasalah
- Ada perbedaan signifikan antar dua ukuran pada Task 9 dan 10 : pengguna kira mereka selesaikan tugas
- Tugas dengan nilai tinggi : Task 3,5 dan 6

RADAR CHART : Five Metrics

Figure 8.4 A sample radar chart summarizing task completion, page visits, accuracy (lack of errors), satisfaction rating, and usefulness rating from a usability test. Each has been transformed to a

Penjelasan Radar Chart

- Lima ukuran : Task completion, page visits, accuracy (lack of error), satisfaction, dan usefulness rating
- Walaupun task completion, accuracy, dan usefulness rating relatif tinggi, page visits dan satisfaction relatif rendah.

HARVEY BALLS

- Dipopulerkan oleh Consumer Reports
- Presentasikan summary data untuk 3 atau lebih ukuran, tapi juga informasikan task-level information
- Contoh : Data dari tabel Z Scores di Tabel 8.7 diubah menjadi Harvey Balls

Tabel Data – SUM

	SUM						
Task	Low	Mean	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%				

Table 8.8 Sample standardized data from a usability test*.
*After entering data for each participant and each task, these are the stand... calculated by SUM, including an overall SUM score and a confidence interval for it.

DIURUTKAN DARI MEAN YANG TERTINGGI

HARVEY BALLS dari Tabel Data SUM

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%				

Excellent 90 – 100%
Very Good 80 – 89%
Good 70 – 79%
Fair 60 – 69%
Poor < 60%

Figure 8.5 A sample comparison chart using data from Table 8.7. 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 coded circles (known as Harvey Balls), as shown in the key.

COMPARISON TO GOALS AND EXPERT PERFORMANCE

Cara untuk menyimpulkan data usability dengan membandingkannya terhadap external standard : PREDEFINED GOAL dan EXPERT PERFORMANCE

COMPARISON TO GOALS

- Hasil usability test dibandingkan dengan GOAL (sasaran, tujuan) yang ditetapkan sebelum tes
- Goal dapat ditetapkan untuk tiap task level(tiap ukuran) atau keseluruhan level

CONTOH TASK SPECIFIC GOAL

- Paling tidak 90% pengguna yang mewakili berhasil pesan kamar hotel yang sesuai
- Membuka akun online baru rata-rata kurang dari 8 menit
- Paling tidak 95% dari pengguna baru mampu beli produk online setelah memilih barang dalam waktu 5 menit

CONTOH OVERALL GOAL

- Pengguna berhasil menyelesaikan paling tidak 90% dari total jumlah tugas
- Pengguna berhasil selesaikan tugas-tugas dalam rata-rata 3 menit per tugas
- Pengguna menilai aplikasi rata-rata 80% menggunakan SUS rating

Goals

- Usability goals diterapkan untuk task completion, time, accuracy dan satisfaction
- Goals harus dapat diukur
- Data yang ditetapkan harus mendukung pencapaian goal

Sample Data of 8 tasks
Target #page visits & Mean actual # page visits

	Target # of Page Visits	Actual # of Page Visits
Task 1	5	7.9
Task 2	8	9.3
Task 3	3	7.3
Task 4	10	11.5
Task 5	4	7
Task 6	6	6.9
Task 7	9	9.8
Task 8	7	10.2

Table 8.9 Sample data from eight tasks showing target number of page visits and mean of actual number of page visits.

CHART ACTUAL VS TARGET

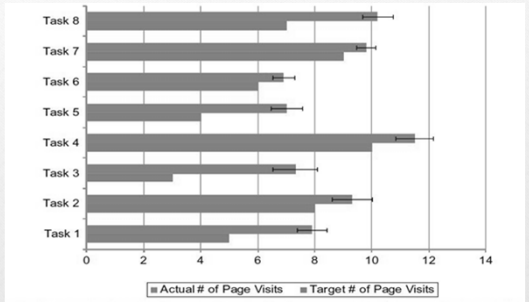


Figure 8.6 Target and actual number of page visits for each of eight tasks. Error bars represent the 90% confidence interval for the actual number of page visits.

Page Visit Efficiency

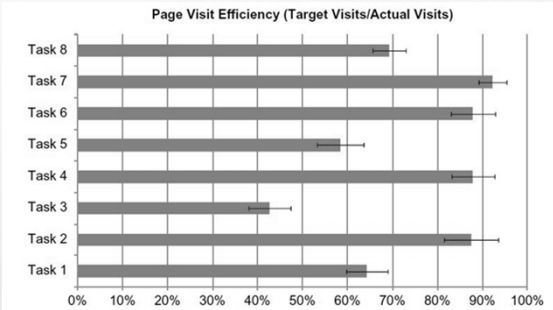


Figure 8.7 Ratio of target to actual page views for each of the eight tasks.

COMPARISON TO EXPERT PERFORMANCE

- Bandingkan hasil usability test dengan performa ahli (expert) terhadap goal/sasaran/tujuan yang ditetapkan
- Beberapa ahli –Subject Matter Expert, menjalani usability test dan hasilnya, dalam bentuk rata-rata, digunakan untuk perbandingan

COMPARISON TO EXPERT PERFORMANCE

- Bermanfaat untuk mengukur tingkat kesulitan beberapa tugas
- Tujuan perbandingan : melihat kedekatan hasil test antara pengguna non expert dan pengguna expert
- Biasanya digunakan pada pengukuran waktu

Sample Data Actual vs Expert Time

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%

Table 8.10 Sample time data from 10 tasks in a usability test showing average actual time per task (in seconds), expert time per task, and ratio of expert to actual time.

Ratio Chart

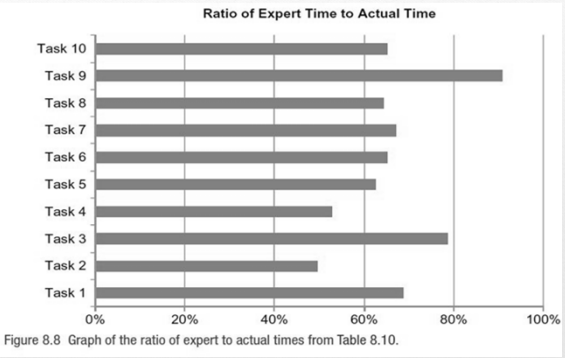
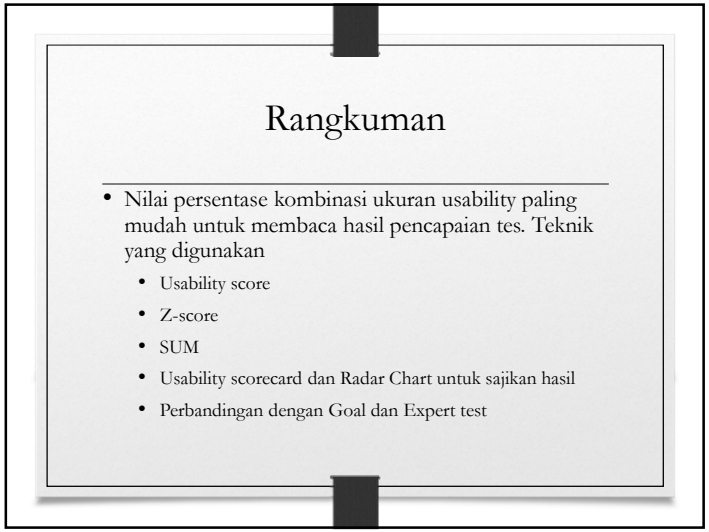


Figure 8.8 Graph of the ratio of expert to actual times from Table 8.10.


Rangkuman
• Nilai persentase kombinasi ukuran usability paling mudah untuk membaca hasil pencapaian tes. Teknik yang digunakan
• Usability score
• Z-score
• SUM
• Usability scorecard dan Radar Chart untuk sajikan hasil
• Perbandingan dengan Goal dan Expert test