Ensuring Web Interface Quality through Usability-based Split Testing
Online Appendix

Table 1. Complete list of interaction features that can be tracked by WaPPU.

label	description	source
arrivalTime	time elapsed from page load till arrival at component	[4]
charsDeleted	# deleted characters	
charsTyped	# characters typed	
clicks	# clicks	[4]
cursorMoveTime	time the mouse cursor spends moving	[4]
cursorRangeX*	cursor range on X axis	[1]
cursorRangeY*	cursor range on Y axis	[1]
cursorSpeed	*cursorTrail* divided by *cursorMoveTime*	[1, 4]
cursorSpeedX	cursor speed in X direction	[1]
cursorSpeedY	cursor speed in Y direction	[1]
cursorStops	# cursor stops	[1]
inputFocusAmount	# focus events on input elements	
cursorTrail	length of cursor trail	[1, 4]
cursorTrailX	length of cursor trail on X axis	[1]
cursorTrailY	length of cursor trail on Y axis	[1]
hovers	# hovers	[4]
hoversPrevHovered	# hovers over previously hovered text elements	[2]
multiplyHoveredText	# multiply hovered text elements	[2]
hoverTime	total time spent hovering the component	[4]
maxHoverTime	maximum time spent hovering the component	[4]
pageDwellTime*	time elapsed between loading and leaving the page	[1]
scrollDirChanges*	# changes in scrolling direction	[3]
scrollMaxY*	maximum scrolling distance from top	[1]
scrollPixelAmount*	total amount of scrolling (in pixels)	[1]
scrollSpeed*	scrolling speed	[1]
textSelections	# text selections	
textSelectionLength	total length of all text selections	

References

1. Guo, Q., Agichtein, E.: Beyond Dwell Time: Estimating Document Relevance from Cursor Movements and other Post-click Searcher Behavior. In: Proc. WWW, pp. 569–578. ACM, New York (2012)
2. Navalpakkam, V., Churchill, E.F.: Mouse Tracking: Measuring and Predicting Users' Experience of Web-based Content. In: Proc. CHI, pp. 2963-2972. ACM, New York (2012)

Table 2. Evaluations by participants familiar with the web search who used an *HD* screen (A = old interface, B = new interface). All tests of significance were carried out as *Mann–Whitney U tests* ($\alpha=0.05$).

usability item	A (N=47)		B (N=42)		significance
	μ	σ	μ	σ	
informativeness	-0.17	0.84	-0.02	0.84	—
understandability	0.34	0.70	0.45	0.67	—
confusion	0.30	0.78	0.38	0.70	—
distraction	0.36	0.74	0.62	0.62	p<0.05, W=798.5
readability	0.45	0.65	0.52	0.71	—
information density	0.04	0.69	0.43	0.67	p<0.01, W=692
accessibility	0.06	0.67	0.07	0.75	—
usability	1.38	2.96	2.45	2.46	p<0.05, W=782

Table 3. Evaluations by participants using an *HD* screen (A = old interface, B = new interface, number in brackets = #samples). All tests of significance were carried out as *Mann–Whitney U tests* ($\alpha=0.05$).

usability item	not familiar					familiar				
	A (47)		B (42)		sig.	A (13)		B (17)		sig.
	μ	σ	μ	σ		μ	σ	μ	σ	
informativeness	-0.17	0.84	-0.02	0.84	—	0.15	0.80	-0.24	0.90	—
understandability	0.34	0.70	0.45	0.67	—	-0.15	0.80	0.41	0.50	—
confusion	0.30	0.78	0.38	0.70	—	0.69	0.48	0.29	0.47	W=171.5 p<0.01
distraction	0.36	0.74	0.62	0.62	W=798.5 p<0.05	0.85	0.38	0.41	0.62	—
readability	0.45	0.65	0.52	0.71	—	0.62	0.51	0.24	0.44	—
information density	0.04	0.69	0.43	0.67	W=692 p<0.01	0.38	0.65	0.41	0.62	—
accessibility	0.06	0.67	0.07	0.75	—	0.38	0.65	0.29	0.47	—
usability	1.38	2.96	2.45	2.46	W=782 p<0.05	2.92	2.25	1.82	2.70	—

3. Nebeling, M., Speicher, M., Norrie, M.C.: W3Touch: Metrics-based Web Page Adaptation for Touch. In: Proc. CHI, pp. 2311–2320. ACM, New York (2013)
4. Speicher, M., Both, A., Gaedke, M.: TellMyRelevance! Predicting the Relevance of Web Search Results from Cursor Interactions. In: Proc. CIKM, pp. 1281–1290. ACM, New York (2013)

Table 4. Evaluations by participants using a *full HD* screen (A = old interface, B = new interface, number in brackets = #samples). All tests of significance were carried out as *Mann–Whitney U tests* (α=0.05).

	not familiar					familiar				
	A (21)		B (31)			A (7)		B (10)		
usability item	μ	σ	μ	σ	sig.	μ	σ	μ	σ	sig.
informativeness	-0.19	0.81	-0.13	0.88	—	0.14	0.38	-0.30	0.95	—
understandability	0.19	0.68	0.10	0.83	—	0.00	0.00	0.40	0.70	—
confusion	0.05	0.74	0.06	0.85	—	0.14	0.90	0.50	0.71	—
distraction	0.14	0.73	0.10	0.87	—	-0.43	0.79	0.80	0.42	W=8 p<0.01
readability	0.05	0.67	-0.06	0.85	—	-1.00	0.00	0.60	0.52	W=0 p<0.01
information density	0.05	0.80	0.19	0.83	—	-0.43	0.53	0.10	0.57	W=19.5 p<0.1
accessibility	-0.05	0.67	-0.19	0.83	—	-0.43	0.53	0.50	0.71	W=11.5 p<0.05
usability	0.24	2.47	0.06	4.55	—	-2.00	2.16	2.6	2.41	W=5.5 p<0.01

Table 5. Quality of models for predicting the usability of the redesigned interface (evaluated using the *RandomForest* classifier provided by the *WEKA Data Mining Software*).

usability item	F-measure	area under ROC
informativeness	0.373	0.601
understandability	0.452	0.588
confusion	0.414	0.502
distraction	0.518	0.632
readability	0.296	0.391
information density	0.325	0.577
accessibility	0.334	0.494