Green: included in the paper
White: not interesting/marginal

ID	Parameters	ID in the PDF	Notes
H1	**Main aspect - Type**	code_measured_____code_sitestype	Overall balance when considering memory, energy, caching, and networking. However, performance-oriented studies have a strong prevalence of real web apps (18/33 studies) wrt synthetic web apps (5/33 studies). Better understanding why this phenomenon is happening is interesting.
H2	**Main aspect - Data analysis**	code_measured_____code_dataanalysis	As expected, descriptive statistics and hypothesis testing are the most used data analysis strategies across all studies main aspects. However, all studies on memory consumption and networking use descriptive statistics only (no effect size estimation, no hypothesis testing). We do not have a clear explanation for this phenomenon though.
H3	**Main aspect - Number of subjects**	code_measured_____code_nrsites	Interestingly, measurement-based experiments on energy consumption tend to involve a lower number of subjects than the other types of experiments (e.g., performance). Indeed, the majority of experiments focussing on energy have less than 10 subjects (8 cases over 19), followed by having less than 50 subjects (5 cases over 19). This result might be seen as an indication of the effort and time required to execute energy-related experiments, which are notoriously more demanding than other types of experiments. Two primary studies on energy are considering a high number of subjects, …
H4	**Main aspect - Tools**	code_measured_____code_tools	No surprises
H5	**Main aspect - Device type**	code_measured_____code_platform	It makes sense that no energy-oriented studies are using emulation. However, there are performance studies which are done on an emulator and it is well-known that emulators are not meant to have a representative performance of the emulated devices.
H6	**Main aspect - Hosting**	code_measured_____code_siteshosted	No surprises
H7	**Main aspect - Network condition**	code_measured_____code_networkcondition	No surprises
H8	**Main aspect - Scope**	code_measured_____code_scope	Usage scenarios have been used more in energy-oriented studies (3/33) rather than the others (1/33 for networking and 1/33 for performance). This result might be explained by the fact that energy consumption is strictly dependent on the total amount of time an operation is performed (E=P*t) and that in order to build a proper assessment of the energy consumed by a web app researchers cannot always rely on a few samples collected in the (usually short) page load time.
H9	**Device type - Browser**	code_platform_____code_browser	No surprises
H10	**Device type - Type**	code_platform_____code_sitestype	No surprises
H11	**Number of subjects - Type**	code_sitestype_____code_nrsites	It does not come as a surprise that experiments with synthetic subjects tend to involve a fewer number of subjects, with 8 cases with less than 10 subjects, 4 having less than 50 subjects, and only 2 experiments having more than 100 studies. Specifically, in Px <e qui si discutono i due paper uno ad uno>. The fact that synthetic subjects are used in experiments with fewer subjects might be an indication of the fact that developing synthetic subjects is time consuming for researchers, who cannot afford to invest time to developing hundreds of synthetic subjects for their experiments.
H12	**Number of subjects - Device Type**	code_platform_____code_nrsites	Emulation-based experiments tend to be used more in experiments with a higher number of subjects. Indeed, differently from experiments with real devices, where we see a prevalence of experiments with less than 10 subjects, experiments with synthetic subjects tend to be more used when more than 50 subjects are considered (<GIVE NUMBERS>). This phenomenon might be an indication that emulation-based

			experiments, which generally last longer, can scale in an easier manner to a higher amount of subjects. Also, among all considered primary studies, there is not a single experiment involving a real device and more than 1000 subjects; this is a confirmation of the intuitive perception that experiments performed on real devices do not scale. In some cases, this limitation can impact the validity of the study, mostly in terms of low statistical power and external validity of the results. Researchers tend to mitigate the potential bias of having a low statistical power by repeating the measures for each trial of their experiment, whereas the bias with respect to the external validity of the experiment is accepted and reported in the discussion of the threats to validity of the considered studies.
H13	**Number of subjects - Tools**	code_tools_____code_nrsites	No surprises
H14	**Number of subjects - Scope**	code_scope_____code_nrsites	Almost all experiments involving the execution of usage scenarios have less than 10 subjects. This result is expected since the execution of usage scenarios tend to take longer than experiments focussing on page load. Interestingly, there are three exception to this observation.
H15	**Type - Network Condition**	code_sitestype_____code_networkcondition	No surprises