Int. J. Human–Computer Studies 195 (2025) 103408

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Understanding preference: A meta-analysis of user studies

Morten Hertzum

Roskilde University, Roskilde, Denmark

ARTICLE INFO

Keywords:
Preference
Workload
NASA-TLX
Task completion time
Error rate
Usability

ABSTRACT

A user's preference for one system over another is probably the most basic user experience (UX) measure, yet user studies often focus on performance and treat preference as supplementary. This meta-analysis of 144 studies shows that while users in general prefer systems with which they achieve lower task time and error rate, they more consistently and more strongly prefer systems that impose lower workload. In only 2 % of the studies a preferred system imposes significantly higher workload than a nonpreferred system. Across the studies, a stronger preference coincides with a larger difference in workload, task time, and error rate. This correlation is strongest for workload, lower for task time, and lowest for error rate. That is, workload is a stronger predictor of preference than performance is, even for the near exclusively utilitarian tasks covered by this meta-analysis. The implications of these findings include that workload should be more fully integrated in research on usability, UX, and design and that it is risky for practitioners to infer preference from performance, or vice versa.

1. Introduction

A user's preference for one system over another is probably the most basic user experience (UX) measure. It has only become more important with the proliferation of systems that are used at the user's discretion rather than mandated by employers or by the absence of alternative options. Discretionary users have the freedom to act on their preference by choosing to adopt the system they prefer, thereby connecting preferences to sales. Yet, preference is often merely a supplementary issue in studies of how users experience and perform with systems. This study investigates the relation between preference and performance.

Nielsen and Levy (1994) investigated the relation between preference and performance on the basis of a meta-analysis. They confirmed the intuition that preference is, in general, positively associated with performance but they also found that in 25 % of the analyzed cases the users preferred the system with which they performed worse. While the former finding is reassuring, the latter is sufficiently intriguing to have its own name: performance-preference dissociation (Andre and Wickens, 1995). It calls for an explanation. This study pursues two possibilities for understanding preference better. First, performance is a multidimensional construct that has, at least, an effectiveness dimension and an efficiency dimension, which can be measured with metrics such as error rate and task time, respectively. Preference may be differentially related to these dimensions (Kiss et al., 2019), thereby suggesting that they should be analyzed separately. Second, the association between performance and preference may be influenced by other factors, such as workload. Performance-workload dissociations exist for both task time and error rate (Hertzum, 2022), thereby suggesting that workload may make an independent contribution to explaining preference. These two candidates for better understanding preference lead to the research question: *To what extent do users prefer systems with which they achieve lower workload, lower task time, and lower error rate?*

To answer this question, we meta-analyze existing studies that report preference, workload, task time, and error rate for pairs of systems. The meta-analysis consists of comparing the degree to which the users in each study preferred one system over the other with the degree to which the workload, task time, and error rate were lower for the preferred system. A meta-analysis is chosen to achieve variety in the users, systems, and tasks included in the study, to quantify the overall trend across these users, systems, and tasks, and to review the association between preference and performance in existing research. Meta-analyses are particularly useful for providing an overview in situations where the existing individual studies yield mixed results, like for performance versus preference. The main contribution of the present meta-analysis is to show the strong correlation between preference and workload. While users in general prefer systems with which they achieve lower task times and error rates, they more consistently and more strongly prefer systems that impose lower workload.

2. Background

The present study is about post-use preferences, that is, preferences

E-mail address: mhz@ruc.dk.

https://doi.org/10.1016/j.ijhcs.2024.103408
Received 27 September 2024; Received in revised form 20 November 2024; Accepted 22 November 2024
Available online 23 November 2024

expressed after the users try a system. Post-use preferences are informed by the experience of performing tasks with the system. In contrast, pre-use preferences are largely expectations and surface-level impressions. Lee and Koubek (2010) find that pre-use and post-use preferences differ. Pre-use preferences are significantly influenced by interface aesthetics and marginally by system usability, whereas post-use preferences are significantly influenced by both aesthetics and usability (Lee and Koubek, 2010).

2.1. Preference, workload, task time, and error rate

Preference is a comparative construct that denotes a greater liking for one alternative over others. In human-computer interaction, it is most commonly measured by asking users who have used two or more systems to rank them by preference, using questions such as "Which interface did you prefer?" (Hornbæk, 2006). It can also be measured by asking users to indicate their preference for each system on a rating scale with endpoints such as "I did not like it" and "I would like to use this system again" (Castillo et al., 2023). Preference is explicitly included in the ISO 9241-210 (2019) definition of UX. The ISO standard first defines UX as "user's perceptions and responses that result from the use and/or anticipated use of a system, product or service" and then notes that "Users' perceptions and responses include the users' emotions, beliefs, *preferences*, perceptions, comfort, behaviours, and accomplishments that occur before, during and after use" (ISO 9241-210, 2019, emphasis added). In classifications of usability measures, preference has been grouped with other self-reported metrics as a measure of satisfaction as opposed to effectiveness and efficiency (Hornbæk, 2006). This grouping implies that preference is a measure distinct from performance metrics.

Workload is about the balance, or imbalance, between the demands imposed by a task, the system used for performing it, and the capabilities of the user performing the task (Hart and Staveland, 1988). That is, workload concerns the human effort necessary to complete the task. A widespread instrument for measuring workload is the Task Load Index (TLX), which consists of six items: mental demand, physical demand, temporal demand, effort, performance, and frustration (Hart and Staveland, 1988). TLX measures self-reported workload and is so widely used that de Winter (2014) states that "workload has become synonymous with the TLX". However, workload can also be measured analytically, physiologically, and by means of a secondary task (Gawron, 2019). A reason for using self-report measures is that experienced workload has genuine consequences: Users who experience their workload as excessive will behave as though they are overloaded, even if workload is low by measures other than self-reporting (Hart and Staveland, 1988). The composite TLX score is mostly expressed on a scale from "Low" (0) to "High" (100) and calculated by taking the mean of the six item ratings. Hart (2006) dubbed this way of calculating the composite score "raw TLX" to distinguish it from the original proposal to include a procedure for assigning more weight to the items most important to the studied task, thereby tailoring the TLX instrument to the task. The argument for omitting the weighting procedure and adopting raw TLX is that the weighting procedure has been found unnecessary because it has little effect on the resulting TLX scores (Byers et al., 1989; Nygren, 1991).

Task time denotes how long it takes for users to complete tasks with a system. It is also referred to as task completion time and time-on-task. Some studies measure clock time as well as perceived time and use the relation among them as an indicator of workload (e.g., Hertzum and Holmegaard, 2013). When task time is used as a performance metric, like in this study, time means clock time and is considered a resource of which successful systems minimize consumption. In some studies, time is measured for parts of a task rather than for the entire task. For example, time until movement onset is a common reaction time metric in studies of pointing devices (e.g., Hertzum and Hornbæk, 2010). Such measures are not measures of task time. In other studies, task time is not reported directly but can be deduced from reported input rates. For example, a standard metric in studies of interfaces for text input is words per minute (WPM) for tasks that consist of having users input a specified text. The reciprocal of this metric (1/WPM) is the time to complete the task, expressed in the unit of minutes per word.

Error rate is a measure of how effectively users complete tasks. It presupposes that tasks have a correct solution, thereby making it possible to distinguish between the tasks that are solved correctly (successes) and those that are not (errors). Many tasks have this property. The ones that do not include the free exploration of a system, the open-ended browsing of social media, and other tasks without a specified goal. In the literature, error rates are calculated on the basis of different metrics for what constitutes the individual success or error (Hornbæk, 2006). One metric is task completion, that is, whether users are able to complete tasks. Another metric is task success, that is, whether users arrive at the correct solution to tasks. Still other metrics concern whether users make errors during the process of completing tasks. For example, the users of a pointing device may click outside the target object, realize the error, reposition the cursor, and click on the target object (Mott and Wobbrock, 2014). Similarly, the driver of a vehicle may drift into another lane, realize the lane excursion, steer the vehicle back in lane, and continue toward the target destination (Kujala, 2013).

2.2. Preference versus performance

As previously mentioned, the overall trend in Nielsen and Levy's (1994) meta-analysis was that users preferred the system with which they performed the best. Performance was measured by task time for some of the 57 studies in the meta-analysis, by error rate for others, and by a metric combining task time and error rate for still others. In keeping with the overall trend, the authors tend to refer to the performance-preference dissociations as "striking counterexamples" or "outliers". However, the Pearson correlation between performance and preference was 0.46. That is, the variation in performance merely explained $0.46^2 = 21$ % of the variation in preference. The percentage explained was somewhat higher in the subset of cases with novice users (38 %) and somewhat lower in the subset with experienced users (13 %), indicating that experienced users' preferences were less sensitive to performance differences. For both novice and experienced users, preference was mainly determined by factors other than performance. In another meta-analysis, Hornbæk and Law (2007) find lower correlations between preference and the performance metrics time and errors. The correlation between preference and time was 0.31 and corresponded to a preferred system being, on average, about 20 % faster than a non-preferred one. The correlation between preference and errors was 0.24, corresponding to mean error rates of about 13 % and 18 % for the preferred and nonpreferred system, respectively. These findings reiterate that performance and preference tend to be positively correlated but they also indicate that the correlations are low to modest and, thereby, that preference is mainly determined by factors other than time and errors.

The two studies above (Hornbæk and Law, 2007; Nielsen and Levy, 1994) do not investigate workload. The association between workload and performance has, instead, been investigated in the meta-analysis by Hertzum (2022), who finds that users generally experience lower workload, as measured by TLX, from systems with which they solve tasks more quickly and with fewer errors. However, the strength of these associations is merely slight to fair and performance-workload dissociations exist for both task time and error rate and for all six TLX subscales. Longo (2018) confirms that workload and usability are somewhat independent constructs and that workload is useful for predicting performance. Because neither Hertzum (2022) nor Longo (2018) investigate preference, they do not contribute to understanding how preferences are influenced by workload. Similarly, Sauro and Lewis (2009) do not investigate preference but find, on the basis of meta-analysis, that task time and error rate correlate less strongly with

post-test than post-task satisfaction (i.e., satisfaction measured after the user has completed all tasks versus each task). Satisfaction is a notion that resembles preference. Probably, preference is most similar to post-test satisfaction because both preference and post-test satisfaction are aggregates across tasks. The correlation between task time and post-test satisfaction was −0.25 and, thereby, of roughly the same magnitude as the correlation between preference and time ($r = 0.31$) in Hornbæk and Law (2007).

Andre and Wickens (1995) argue that dissociations between preference and performance are unsurprising because "key features that may influence preference (such as aesthetics, novelty, familiarity, or low effort) are not necessarily the same ones that result in effective performance." For example, Schenkman and Jönsson (2000) find that aesthetics, in terms of the beauty of a website, is important to whether users prefer it to other websites. Liu et al. (2020) find that users prefer novel designs to less novel ones for hedonic products and, in the case of users more sensitive to gains than losses, also for utilitarian products. Backhaus et al. (2018) find that elderly users, but not younger users, prefer a smartphone design with familiar metaphors to one with a minimalistic flat design. Chen et al. (2016) find a preference for a system that imposed lower workload on its users even though they were more accurate with another system for controlling the needle during suturing in minimally invasive surgery. An additional feature that has been found to influence preference is that users opt not to exclude themselves from functionality. Frøkjær et al. (2000) compared the full version of an information retrieval system with three restricted versions that contained different subsets of the functionality of the full version. The users overwhelmingly preferred the full version but performed tasks faster with two of the restricted versions.

The existence of performance-preference dissociations leads Andre and Wickens (1995) to the conclusion that, on their own, preference ratings may be misleading. Therefore, they recommend that "performance measures should *always* augment preference ratings" (Andre and Wickens, 1995, emphasis in original). This recommendation is consistent with Bailey (1993), who is concerned that companies may too often make design decisions on the basis of users' preference ratings. In contrast, Nielsen and Levy (1994) propose that preference ratings, which are easy to collect, can replace more costly performance measurements without incurring a high risk of faulty conclusions about which is the better of competing designs. These contrasting opinions demonstrate the importance of understanding the extent to which performance and preference are associated.

3. Method

A systematic procedure was followed to identify and analyze a set of user studies that empirically compared two digital systems with respect to preference, workload, task time, and error rate. This procedure adhered to standard recommendations for systematic reviews and meta-analysis (Littell et al., 2008) and consisted of formulating inclusion criteria, inspecting a total of 8183 papers for inclusion or exclusion, and meta-analyzing the contents of the 144 included papers. Note that System Usability Scale (SUS) scores (Brooke, 1996) were initially included as a measure of preference, and Subjective Workload Assessment Technique (SWAT) scores (Reid and Nygren, 1988) as a measure of workload. Thus, they appear in the list of the inclusion criteria in Section 3.1. However, SUS and SWAT were subsequently excluded from this meta-analysis, see Section 3.2.

3.1. Inclusion criteria

The selection of papers for inclusion in this study was governed by seven criteria formulated prior to the selection process. To be included, a paper had to meet all seven criteria. First, papers had to report preference as measured by preference ratio, ranked preference, preference score, or SUS score. Second, papers had to report workload as measured by TLX or SWAT. Third, papers had to report task time. Fourth, papers had to report task completion as measured by error rate, task completion rate, or task success rate. Fifth, papers had to be empirical studies comparing the use of digital systems or the use of a digital system and a manual system. Sixth, papers had to be peer-reviewed research published in journals, at conferences, or as book chapters. Seventh, papers had to be in English.

3.2. Selection process

The process of selecting the papers for inclusion in this study involved multiple steps, see Fig. 1. First, four databases were searched for papers that contained terms suggesting coverage of the first four inclusion criteria. With small syntactic variations across the databases, the search query was: (preference OR satisfaction OR "system usability scale" OR sus) AND ("task load index" OR tlx OR "subjective workload assessment technique" OR swat) AND time AND ("error rate" OR "task completion" OR "task success"). The four databases were ACM Digital Library, Google Scholar, Scopus, and IEEE Xplore. They were chosen for their coverage of potentially relevant papers and searched in January 2024. Second, duplicates were removed from the set of 8183 papers that matched the query, and then the unique papers were screened. The papers were screened by matching their title against the inclusion criteria and, if they passed this screening, by matching their abstract against the inclusion criteria. After these two screenings, 1181 papers remained. Third, the full text of these papers was looked up. All but five of them were obtainable online. The five unobtainable papers were requested from their authors, who in three cases kindly supplied a full-text copy. Fourth, the full text of the 1179 obtained papers was matched against the inclusion criteria. The main reasons for excluding papers at this stage were that they did not report data about preference or task completion (Fig. 1). After the full-text screening, 222 papers remained. Fifth, while these 222 papers met the inclusion criteria, it was post hoc decided to exclude 78 of them. To sharpen the focus on preference, the papers that reported a SUS score rather than a genuine measure of preference were excluded (SUS includes an item about preference but its other items are more broadly about usability). With these exclusions, preference was always the users' indication of their preferred system. In addition, only 1 of the 222 papers measured workload with SWAT. This paper was excluded to achieve a uniform conceptualization of workload. With this exclusion, all 144 included papers measured workload with TLX.

3.3. Data analysis

The analysis of the 144 included papers proceeded in six steps. First, many of the papers compared more than two systems. If all pairwise comparisons in these papers were included in the meta-analysis, it would become biased (Littell et al., 2008). For example, Jeon et al. (2009) compared five systems and would, thus, contribute ten pairwise comparisons involving the same users and study setup. To avoid such bias, one pairwise comparison was selected from each included paper, namely the comparison between the two systems with the largest difference in preference. The rationale for this selection was that it is more interesting to understand how preference relates to workload, task time, and error rate when the difference in preference is large than when it is small and, possibly, inconsequential or down to chance.

Second, data about preference, workload, task time, and error rate were extracted for each of the 144 pairwise comparisons. The papers reported *preference* in four different ways: (1) the number of users preferring each system, (2) the percentage of users preferring each system, (3) the users' rating of each system on a preference scale, and (4) the users' rank-ordering of the systems. *Workload* was predominantly reported as raw TLX scores. Only a few studies (e.g., Hu and Malthaner, 2007) also employed the weighting procedure, which tailors the TLX instrument to the studied task. Because the weighting procedure has

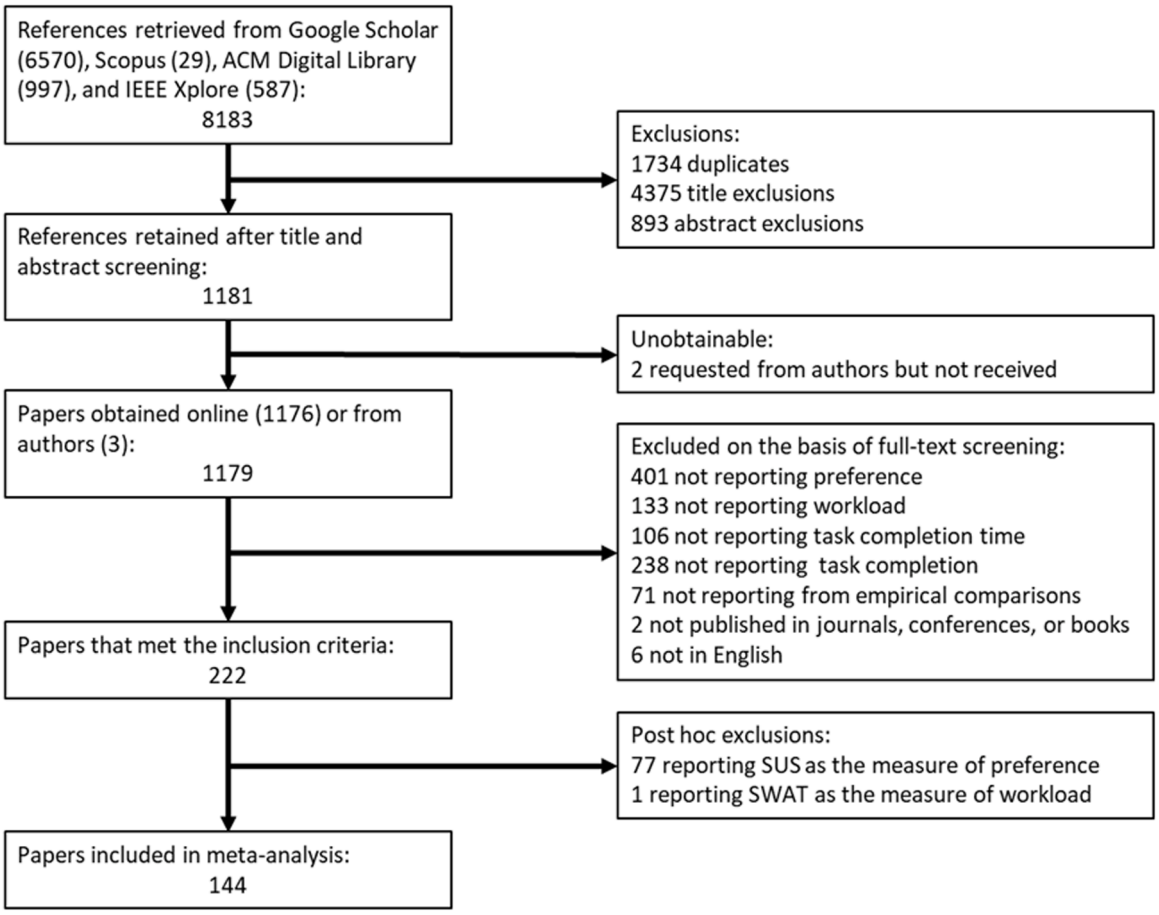

Fig. 1. Paper-selection process

been found to have little effect on the resulting TLX scores (Byers et al., 1989; Nygren, 1991), this meta-analysis did not distinguish between raw and weighted TLX. *Task time* and *error rate* were the main performance measures in most of the studies and often reported in more detail than preference and workload. The extraction process involved copying data that were directly available and, in some cases, reading values from graphs or calculating needed data from the data available. For example, data about overall workload (i.e., raw TLX) were in several cases not provided directly but could be calculated by averaging available data about the six TLX subscales. Relatedly, input rates expressed in WPM were converted to task times, and success rates were converted to error rates.

Third, the extracted data for two of the measures were scores on rating scales: the preference ratings and TLX scores. These data were rescaled to the interval from 0 to 1 and arcsine transformed. The arcsine transformation stretched out values close to the end point of the scales, thereby compensating for their fixed end points, which make it harder for a mean rating to get close to the ends of the scale. Values were rescaled using the formula: (value − lower endpoint) / (max endpoint − lower endpoint). This formula corresponded to the one used by Lewis and Erdinc (2017).

Fourth, the data were analyzed by simply counting the number of studies where users preferred the system with which they achieved lower workload, lower task time, and lower error rate. This analysis was, however, coarse-grained. To quantify the magnitude of the difference between the two systems, an effect size was also calculated. Many papers provided means but no information about the distribution (e.g., no standard deviations), thereby limiting the choice of effect-size measures for the meta-analysis. Like in Nielsen and Levy (1994), the chosen measure was the logarithm of the ratio of the means: log(mean for system 1 / mean for system 2). For example, if the mean task time was 20 seconds for system 1 and 10 seconds for system 2, then the effect size would be $\log(20/10) = 0.301$. Conversely, if the mean task time was 10 seconds for system 1 and 20 seconds for system 2, then the effect size would be $\log(10/20) = -0.301$. The same difference in performance – in this case that one system is twice as fast as the other – results in the same absolute effect size irrespective of whether the faster system is system 1 or system 2; the only difference is the sign of the effect size. This feature of the logarithm function was the rationale for log-transforming the ratios. While an effect size could be calculated for almost all the data, it could not be calculated for the preference rankings because they merely stated which system the users preferred and, thus, gave no information about the magnitude of the preference. Therefore, the 13 studies in which preference was measured by preference rankings had to be excluded from this part of the meta-analysis, leaving 131 studies.

Fifth, the effect-size measure was only defined when the values for system 1 and 2 were non-zero. Values of zero occurred in two ways and were replaced with imputed values. In studies with unanimous preference for one of the two systems, zero users preferred the other system. In these cases, the preference for the nonpreferred system was set to 1 divided by the square root of the number of users in the study. For example, in a 16-user study, it would be set to $1/\text{sqrt}(16) = 0.25$ users. This formula always set the imputed number of users to less than one, and it incorporated that unanimous preference was harder to achieve when more users participated in a study. The second way in which values of zero occurred was when error rates equaled zero. In these cases, the error rate was set to $1/\text{sqrt}(n)$ percent, where n was the number of users in the study.

Sixth, the studies were characterized by extracting and categorizing information about the involved users, tasks, systems, domains, and the

countries in which the studies were conducted. The user categorization distinguished between novices, who had no or very little experience with the tasks and systems, and experienced users, who had some experience though they in most cases could not be considered experts. The task categorization distinguished between utilitarian tasks, which were about effectively and efficiently achieving practical goals, and hedonic tasks, which were about having enjoyable or playful experiences. The system categorization distinguished among five categories of systems, see Table 1. Finally, the domain categorization distinguished between safety-critical domains and domains that were not safety critical.

4. Results

The 144 included studies spanned both novice and experienced users and a variety of systems in different categories, see Table 1. The table also shows that the tasks were near exclusively utilitarian, that most of the studied domains were not safety critical, and that the majority of the studies were conducted in Europe and North America. In addition, all but one study had a within-subjects design in which the users performed tasks with both systems in the pairwise comparison before they indicated their preference. In the last study, each user performed tasks with only one of the two systems and then expressed their preference on a rating scale. A total of 2716 users participated in the studies, corresponding to a mean of 18.9 participants per study.

4.1. Counts of dissociation with preference

In 86 (60 %) of the 144 studies, the users preferred the system with which they achieved lower workload, lower task time, and lower error rate, see Table 2. That is, the data confirmed the intuition that users can in general be expected to prefer performing well and to use systems that help them rather than hinder them. However, the remaining 58 (40 %) of the studies displayed a dissociation between preference and one or several of workload, task time, and error rate. Most of these dissociations were the result of disagreement among the three latter measures. For example, Turner et al. (2021) compared interfaces for smartwatches and found that users were faster and experienced lower workload with a trace-based interface but made fewer errors with a tap-based interface. In these cases, the users' preference indicated which of workload, task time, and error rate they considered more important when a preference for neither one nor the other system would give them lower workload, lower task time, as well as lower error rate. The users in the study by Turner et al. (2021) preferred the trace-based interface and thus displayed a dissociation between preference and error rate.

Notably, the users in three studies (Mathis et al., 2021; Prilla and Mantel, 2021; Ranasinghe et al., 2019) preferred the system with which they experienced higher workload, higher task time, as well as higher error rate (Table 2). For example, Mathis et al. (2021) studied mechanisms for obtaining usable and secure authentication in virtual reality. The users preferred an authentication mechanism operated by means of eye gazes, but they experienced lower workload, task time, and error rate with an authentication mechanism operated by tapping on a physical controller. Mathis et al. (2021) explained this threefold dissociation by interpreting the users' preference mainly as a preference for high security and their performance with the preferred authentication mechanism as an indication of poor usability. This explanation of the dissociation is consistent with the established tension between usability and security in studies of authentication mechanisms. In the two other studies with a threefold dissociation, the authors appeared to assign primacy to preference. Prilla and Mantel (2021) saw the threefold dissociation in their study as a reminder that good design is about more than low workload, task time, and error rate and suggested that systems must also be customizable. Ranasinghe et al. (2019) deemphasized the results about workload, task time, and error rate by noting that they were not statistically significant, thereby leaving the difference in preference as the main result of their study.

Table 3 aggregates the information in Table 2 to show the total number of associations and dissociations for workload, task time, and error rate. The number of dissociations was lowest for workload (9 %), thereby indicating that it was the aspect most important to users' preference. In addition, the number of dissociations was slightly lower for task time (19 %) than error rate (24 %). The users preferred the system with higher error rate in nearly one of every four studies. However, the difference in workload, task time, and error rate between the preferred and nonpreferred system varied from small in some studies to large in

Table 1
Profile of the included studies, $N = 144$ studies

Category	Examples	Count	Percentage
Users			
Novices	First-time users, Non-target users, Students	58	40
Experienced	Long-term computer users, Nurses, Pilots	68	47
Other	Mixed user group, Unspecified user group	18	13
Tasks			
Utilitarian	Text entry, Object selection, Menu navigation	140	97
Hedonic	Assembling Lego model, Playing computer game	4	3
Systems			
Graphical user interfaces	Smartwatch, Website, In-vehicle information system	67	47
Augmented reality	Assembly assistance, Order picking, Care provision	26	18
Virtual worlds	Gaze interaction, Teleportation, Text entry	32	22
Teleoperation	Surgical equipment, Space robots, Assembly tasks	6	4
Non-visual interfaces	Voice input, Gaze authentication, Prothesis control	13	9
Domains			
Safety critical	Driving, Healthcare, User authentication	30	21
Not safety critical	Video navigation, Text analysis, One-handed input	114	79
Regions			
Europe	Germany, UK, Austria	60	42
North America	US, Canada	54	38
Asia	China, Republic of Korea, Japan	25	17
Australasia	New Zealand	3	2
Africa	Egypt	1	1
South America	Brazil	1	1

Table 2
Frequency of all combinations of association/dissociation with preference, $N = 144$ studies

Workload	Task time	Error rate	Count	Percentage
Association	Association	Association	86	60
Association	Association	Dissociation	24	17
Association	Dissociation	Association	16	11
Association	Dissociation	Dissociation	5	3
Dissociation	Association	Association	3	2
Dissociation	Association	Dissociation	3	2
Dissociation	Dissociation	Association	4	3
Dissociation	Dissociation	Dissociation	3	2

Table 3
Frequency of association/dissociation with preference

	Association		Dissociation		Total	
	N	%	*N*	%	*N*	%
Workload	131	91	13	9	144	100
Task time	116	81	28	19	144	100
Error rate	109	76	35	24	144	100

others. For a number of the dissociations, the users' higher workload, task time, or error rate with the preferred system did not reach statistical significance. That is, the users in these studies performed worse, but not significantly worse, with the preferred system.

Tests of statistical significance were available in only a subset of the 144 studies (e.g., because several of the studies that compared more than two systems did not report statistical tests for all pairwise comparisons). Table 4 is restricted to these studies; it shows the number of associations and dissociations for the subset of studies that reported whether the difference in workload, task time, or error rate was statistically significant. The users preferred the system that imposed significantly lower workload in 63 % of the studies and the system that imposed significantly higher workload in 2 % of the studies; in the remaining 35 % of the studies the difference in workload was not statistically significant. The number of dissociations was about four times higher for task time (8 %) and error rate (7 %) than for workload. That is, the users occasionally preferred a system with significantly higher workload, task time, or error rate and when they did, they more often preferred longer task time or higher error rate than higher workload.

4.2. Preference versus workload

To quantify the magnitude of the difference between the two systems in each study, the logarithm of the ratio between them was calculated for preference, workload, task time, and error rate. Fig. 2 shows the result for preference and workload by plotting them against each other. Positive values on the horizontal axis indicate a preference for the first system in a study and negative values a preference for the second system (note that it has no impact on the outcome of the analysis whether a system is designated as the first or second system because the result would be the same whether first was compared to second or second to first). Similarly, positive values on the vertical axis indicate that the first system imposed higher workload and negative values that the second system imposed higher workload. That is, the users in the studies plotted in the upper-left and lower-right quadrants preferred the system that imposed the lower workload. The studies in the two other quadrants (lower-left and upper-right) displayed a preference-workload dissociation. These studies tended to have a small difference in preference, workload, or both.

The overall trend in the data was that with increasing difference in preference, the workload imposed by the preferred system was increasingly lower than that imposed by the nonpreferred system. This trend was confirmed by linear regression; the variation in preference explained 54 % of the variation in workload. For example, Özacar et al. (2016) compared a flick technique for selecting objects in an augmented reality system by finger movements with a head cursor for making the selections by head movements. The users experienced that the flick technique, which was preferred by zero users, imposed a workload of TLX = 49, whereas the head cursor, which was preferred by eight users, imposed a workload of TLX = 18. That is, the unanimous preference for the head cursor coincided with nearly three times lower workload. Relatedly, Kytö et al. (2018) found that a smaller difference in preference coincided with a smaller difference in workload. In their study, 58 % of the users preferred a head-interaction technique for selecting

Table 4
Frequency of association/dissociation with preference for the subset of studies that reported whether the difference in workload, task time, and error rate was statistically significant

	Association and sign. difference		No sign. difference		Dissociation and sign. difference		Total	
	N	%	*N*	%	*N*	%	*N*	%
Workload	55	63	31	35	2	2	88	100
Task time	74	64	32	28	9	8	115	100
Error rate	43	40	57	53	8	7	108	100

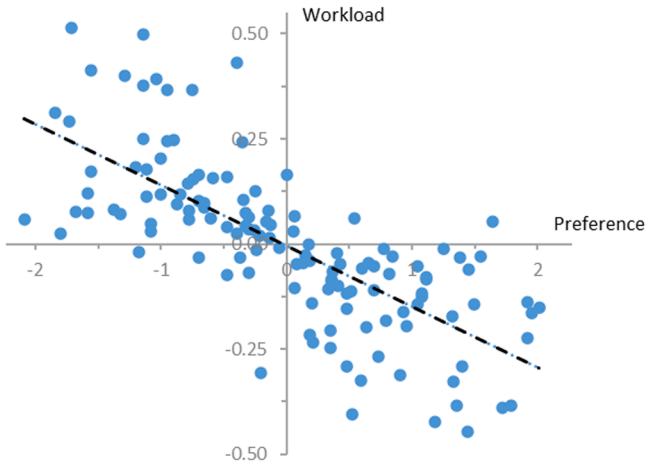

Fig. 2. Scatterplot of preference versus workload, $N = 131$ studies. Each point represents a study comparing two systems and shows the logarithm of the ratio of the preferences plotted against the logarithm of the ratio of the workload scores. The dashed trendline was determined by linear regression, $y = -0.145x - 0.004$ ($R^2 = 54$ %).

objects in augmented reality, whereas 42 % preferred an eye-interaction technique for making the selections. This modest difference in preference coincided with a small 36 (head interaction) versus 42 (eye interaction) difference in TLX.

4.3. Preference versus task time

Fig. 3 plots preference against task time. Save a single outlier, the values for task time were roughly in the same range as those for workload (time range: −1.05 to 0.60, workload range: −0.45 to 0.51). Like for workload, the overall trend in the data was that increased preference coincided with larger performance improvements: With increasing difference in preference, the task time with the preferred system was increasingly lower than that with the nonpreferred system. However, the association between preference and task time was weaker than that between preference and workload. The variation in preference explained a smaller percentage ($R^2 = 36$ %) of the variation in task time than in workload. This finding was in accord with the larger number of dissociations between preference and task time. For example, Jardina and Chaparro (2012) compared the Nook and Kindle e-readers on

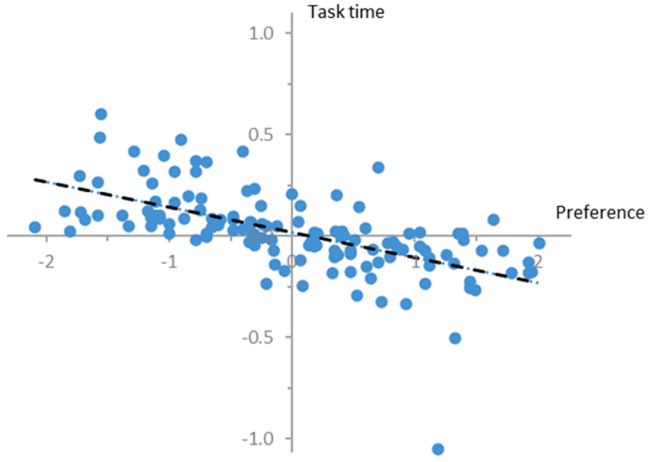

Fig. 3. Scatterplot of preference versus task time, $N = 131$ studies. Each point represents a study comparing two systems and shows the logarithm of the ratio of the preferences plotted against the logarithm of the ratio of the task times. The dashed trendline was determined by linear regression, $y = -0.124x + 0.016$ ($R^2 = 36$ %).

different book-navigation tasks. The physical dimensions of both e-readers were 155 × 91 mm. Ten users preferred the Nook although they completed tasks 2.19 times faster (70s vs 32s) with the Kindle, which was preferred by only two users. That is, a fivefold preference for the Nook was contradicted by sizably longer task times. The preference was possibly explained by a 0.18/0.37 = 0.49 times lower error rate for the Nook.

4.4. Preference versus error rate

Fig. 4 plots preference against error rate. The values for error rate were more dispersed (range: −2.52 to 2.58) than for workload and task time because the difference in error rate between the two systems was extensive in some studies. For example, Kim et al. (2014) compared different sizes of touch keys in an in-vehicle information system (IVIS) and found an error rate of 1 % for the preferred key size of 22.5 mm and 15 % for the nonpreferred key size of 7.5 mm. This 15-fold difference coincided with a similarly large difference in preference. The keys received mean preference ratings of 75.11 (22.5 mm keys) and 6.13 (7.5 mm keys) on a 0-100 scale, a 12.3-fold difference. In contrast, Schramm et al. (2023) found unanimous preference for a gaze-controlled IVIS with a 22-fold higher error rate than a gesture-controlled IVIS, probably because the error rate was low for both systems (eye gaze: 2.86 % vs hand gesture: 0.13 %). In spite of frequent and sometimes large dissociations between preference and error rate, the overall trend in the data was that with increasing difference in preference, the error rate with the preferred system was increasingly lower than that with the nonpreferred system. However, the variation in preference explained merely 19 % of the variation in error rate.

4.5. Predicting preference from performance

A multi-regression model with workload, task time, and error rate as predictors significantly predicted preference, $F(3, 127) = 56.09$, $p < 0.001$. The values entered into the model were those depicted in Figs. 2-4. The resulting standardized coefficients (β) allowed for comparing the strengths of the predictors. Workload was the strongest predictor (β = −0.551), followed by task time (β = −0.195) and error rate (β = −0.112). The relative strength of the predictors was that workload had a −0.551/−0.195 = 2.8 times stronger effect on preference than task time had and a −0.551/−0.112 = 4.9 times stronger effect than error rate had. Collectively, the variation in workload, task time, and error rate explained 57 % of the variation in preference. Because of co-variation between workload, task time, and error rate, this percentage was only marginally higher than the 54 % explained by the variation in workload alone (Fig. 2).

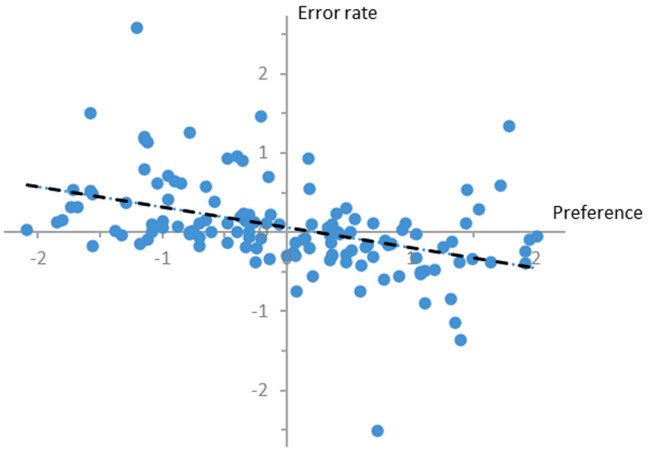

Fig. 4. Scatterplot of preference versus error rate, $N = 131$ studies. Each point represents a study comparing two systems and shows the logarithm of the ratio of the preferences plotted against the logarithm of the ratio of the error rates. The dashed trendline was determined by linear regression, $y = -0.257x + 0.056$ ($R^2 = 19$ %).

4.6. Subset analysis

The categorization of the included studies into different categories of users, domains, and so forth (Table 1) made it possible to analyze how strongly preference correlated with workload, task time, and error rate for the studies in the different categories. Such analyses were made for the categories of users and domains, see Table 5. The values entered into these analyses were subsets of the values depicted in Figs. 2-4. Overall, preference correlated strongly with workload and task time and moderately with error rate. However, these correlations for the set of 131 studies masked noteworthy differences between novice and experienced users and between safety-critical and non-safety-critical domains.

Workload, task time, and error rate correlated more strongly with preference for experienced than novice users. That is, experienced users were more likely than novices to prefer the system with which they experienced lower workload, task time, and error rate. This difference between novices and experienced users was most pronounced for workload and task time. Specifically, the variation in workload explained 20 percentage points more of the variation in preference for experienced ($R^2 = 67$ %) than novice ($R^2 = 47$ %) users, the variation in task time explained 15 percentage points more of the variation in preference (45 % vs 30 %), and the variation in error rate explained 7 percentage points more of the variation in preference (24 % vs 17 %).

With respect to domain, preference correlated less strongly with workload in safety-critical domains, suggesting that users in these domains accepted a higher workload to maintain a high level of safety, whereas users in domains that were not safety critical had less reason to prefer a system that imposed high workload. The variation in workload explained 12 percentage points less of the variation in preference for safety-critical (46 %) than non-safety-critical (58 %) domains. In contrast, the percentage explained by task time and error rate was near identical for safety-critical and non-safety-critical domains.

5. Discussion

Preference is often measured in user studies, but they rarely assign primacy to it. This meta-analysis investigates the relations between preference, workload, task time, and error rate for a variety of systems with the common feature that they near exclusively are for utilitarian tasks.

5.1. Main findings

This study contributes five main findings. In the following, these findings are summarized and compared with those of the two previous meta-analyses that relate preference to performance, but not to workload (Hornbæk and Law, 2007; Nielsen and Levy, 1994).

Table 5
Correlation with preference for subsets of the studies

Category	*N*	Workload	Task time	Error rate
All	131	−0.734***	−0.604***	−0.431***
Users				
Novice	49	−0.686***	−0.544***	−0.413**
Experienced	65	−0.818***	−0.669***	−0.489***
Domain				
Safety critical	26	−0.681***	−0.639***	−0.458*
Not safety critical	105	−0.761***	−0.615***	−0.428***

* $p < 0.05$,
** $p < 0.01$,
*** $p < 0.001$ (Pearson correlation)

First, *preference is in general associated with workload, task time, and error rate*. In 60 % of the 144 included studies, the users prefer the system with which they achieve lower workload, lower task time, and lower error rate. In an additional 30 % of the studies, the users prefer the system with which they achieve two but not all three of lower workload, task time, and error rate. This finding corresponds roughly with Nielsen and Levy (1994), who found agreement between performance and preference in 75 % of the studies they reviewed. It also corresponds with what Hornbæk and Law (2007) call the half-full interpretation of their data. This interpretation emphasizes – with some surprise – that preference in most cases correlates positively with performance, though Hornbæk and Law (2007) find lower mean correlations than the present study.

Second, *workload is a stronger predictor of preference than performance is*. The variation in workload explains 54 % of the variation in preference, while the variation in task time explains 36 % and the variation in error rate 19 %. Multi-regression showed that workload has a 2.8 and 4.9 times stronger effect on preference than task time and error rate, respectively. Furthermore, dissociations between preference and workload tend to be restricted to studies with a small difference in preference, workload, or both. Workload is significantly higher for the preferred system in only 2 % of studies (Table 4). The strong association between workload and preference indicates that workload is a variable important to users and, therefore, that it should be included in system evaluations. Hornbæk (2006) finds that this has mostly not been the case in the past, only 5 % of the studies in his review included workload measurements.

Third, *task time is a stronger predictor of preference than error rate is*. This finding shows that the two most common measures of performance interact with preference in different ways, thereby elaborating Nielsen and Levy's (1994) finding of a 0.46 correlation between preference and performance. The present meta-analysis finds a similar correlation between preference and error rate (−0.43) but a stronger correlation between preference and task time (−0.60). This difference means that the variation in task time explains 17 percentage points more of the variation in preference than error rate does. The finding of a stronger correlation for task time than error rate accords with Hornbæk and Law (2007) though they report lower mean correlations between preference and task time as well as between preference and error rate. That is, the present study finds users more likely to prefer the system with which they perform best.

Fourth, *preference is occasionally dissociated from workload, task time, or error rate*. In four of every ten studies, the users prefer the system with which workload, task time, or error rate is higher without necessarily being significantly higher. Looking only at the studies that report tests of statistical significance, workload is significantly higher for the preferred system in 2 % of studies, task time is significantly longer for the preferred system in 8 % of studies, and error rate significantly higher for the preferred system in 7 % of studies. That is, performance-preference dissociations are too frequent to be explained away as outliers, as Nielsen and Levy (1994) tend to do. Consistent with Hornbæk and Law's (2007) half-empty interpretation of their data, performance-preference dissociations are so common that we would be ill-informed if their presence surprised us. They are several times more frequent than preference-workload dissociations.

Fifth, *experience level and safety criticality moderate the findings*. For experienced as opposed to novice users, more of the variation in preference is explained by the variation in workload (67 % vs 47 %), task time (45 % vs 30 %), and error rate (24 % vs 17 %). That is, experienced users' preferences are more sensitive to workload and performance. Possibly, novices relate their workload and performance to their process of learning to use a system and consider it a somewhat unrelated issue whether the system will be preferable once learned. This finding contradicts Nielsen and Levy (1994), who find that performance correlates more strongly with preference for novice than experienced users. For safety criticality, the variation in workload explains less of the variation in preference for safety-critical than non-safety-critical domains (46 % vs 58 %). This finding indicates that users are less concerned with workload when safety is at stake.

5.2. Preference formation

Understanding how preferences are formed involves explaining "how beliefs and evaluations emerge from correlations between what people experience and what they feel" (Druckman and Lupia, 2000). These beliefs and evaluations can be influenced by social processes, such as product advertisements, organizational norms, and peer pressure. However, the present study is restricted to how the concrete experience from using a system influences the user's preference for or against the system. Systems have multiple features, each contributing to this experience in ways that may vary across users, tasks, and situations. A preference is the individual user's integration of these multiple orientations toward the system into an overall evaluation. Two models attempt to describe the process by which the user makes this integration (Druckman and Lupia, 2000):

The memory-based model asserts that users remember their experiences with the system and their multiple orientations toward it. They form their preferences by retrieving information about these experiences and orientations from memory. The information retrieved is merely a subset of the information held in memory about how the user has experienced the system. The likelihood of retrieving a piece of information increases with its accessibility, which in turn increases with the recency, frequency, and consonance of the information (Matthes, 2007). As a result, preferences are susceptible to sudden shifts, but these shifts may reflect differences in the retrospectively attended information to a larger extent than differences in the user's moment-to-moment experience of a system. The risk of such instability also exists in studies like those in this meta-analysis because the users only experienced the systems for a limited set of tasks in a lab-like setting that necessarily lacked the richness of everyday system use. That said, the users expressed their preference immediately after using the systems, that is, with their experience of using the systems still fresh in their mind.

In contrast to the memory-based model, *the on-line model* asserts that users form their preference while they are exposed to the systems rather than when they, subsequently, need to report or act on it. To explain this model, it can be imagined that users merely keep a mental tally of their orientation toward a system. When they have new experiences with the system, they retrieve the tally from memory, update it, and return it to memory. After updating the tally, the user may, and often will, forget the experience that caused the update. When asked for their preference, users simply retrieve the tally. By retrieving the tally, they know how strongly they prefer, or disprefer, a system. Importantly, they are often unable to substantiate their preference with information about the specific experiences on which the tally is based because this information has not been remembered. This model lends itself well to asking the users to indicate their preference by ranking or rating the tested systems, rather than by interviewing the users about how they experienced the systems. Because preferences according to the on-line model are available in memory as tallies, this model suggests that preferences are more stable than the memory-based model would imply (Druckman and Lupia, 2000).

Matthes (2007) contends that both models are required to understand preference formation. When users initially encounter a system, no previously formed preference is available in memory. Hence, they engage in forming a preference on the basis of accessible information about their experience with the system. Later, this memory-based preference may crystallize into an on-line preference if the users gain more experience with the system and if these experiences tend to be consonant. Once crystallization has happened, preferences tend to remain stable over time, partly because "any new information will be biased in the direction of the initial on-line judgment" (Matthes, 2007). Some studies indicate that crystallization happens quite quickly. For example, Lindgaard et al. (2006) find that people form an opinion about the visual

appeal of a website well within the first second of being exposed to it. On-line preferences can also revert back to memory-based ones. Such attenuation is most likely to happen when there is a decrease in users' need for having a ready opinion about whether they prefer one system or another (Matthes, 2007).

A key finding of this study is that preference correlates strongly with workload. This finding accords with Quinn and Cockburn (2020) who find that users are loss averse and prefer systems with which they do not lose work even when these systems do not lead to lower task time and fewer errors. That is, they prefer lower workload for its own sake, not just when it coincides with better performance. Relatedly, users prefer systems that lead to small but reliable progress over alternative systems that are susceptible to all-or-nothing outcomes even if these alternative systems are consistently faster when successful. That is, users prefer systems that impose a predictable workload to systems that impose low or high workload depending, partly, on unpredictable situational factors. Evidence for this sensitivity to workload comes from studies finding that users prefer finding files by navigating to them rather than searching for them (Bergman et al., 2008), prefer clicking their way through menus to remembering and using keyboard shortcuts (Lane et al., 2005), and prefer repeatedly using a small set of general-purpose commands to learning and adopting a large set of commands made for specific purposes (Thomas, 1998). These studies also indicate that users may be more sensitive to their immediate workload than to long-term workload.

5.3. Implications

Preferences matter. At the level of reasoning in general, people are considered rational if they act in accordance with their own preferences – and moral if they act in accordance with everybody's preferences (Fehige and Wessels, 1998). At the level of UX research and practice, preferred prototypes are more likely to be developed into final designs, preferred products are more likely to be purchased, and – as this study shows – preferred systems are more likely than not to impose lower workload and better performance.

For practitioners, the overarching implication of this study is to recognize that performance and preference are related but at the same time somewhat independent, even for utilitarian tasks. The independence is probably more pronounced for hedonic tasks, but conclusions about hedonic tasks are beyond the scope of this study. Several more specific implications follow from the overarching implication. First, preference should be included in system evaluations because it is a basic UX metric and easy to measure. If several systems are evaluated, users can simply indicate which one they prefer. If only a single system is evaluated, users can indicate whether they would prefer it to their current ways of working. In discount evaluations, it may be tempting to bypass performance measurements to expedite evaluations, thereby using preference as a proxy for performance. Fig. 2-4 can be used to estimate whether the resulting difference in preference is likely to correspond to lower workload, task time, and error rate. Second, workload should be included in system evaluations because it correlates strongly with preference and has diagnostic power beyond that of task time and error rate. It may specifically be noted that workload and task time are both measures of efficiency but capture different aspects of it. The TLX instrument (Hart and Staveland, 1988) makes workload easy to measure, and Hertzum (2021) provides reference values for TLX and its subscales to help interpret the measurements. Third, persuasion may be needed to convince users that the best performing system should be preferred. To become aware of the need for persuasion, both performance and preference must be measured. In the case of dissociation, the first step should be to understand the reasons for the preference and to weigh them against the importance of better performance. Many companies see lower task time as a means of reducing expenses and increasing profits. Companies in safety-critical domains are particularly concerned with avoiding errors because they may have severe consequences. After understanding the reasons, the next step can either be to change the system to reconcile the dissociation or, if such changes are not possible, to work with the users to make them aware of its consequences to the company.

For researchers, the main implication of this study is to recognize that preference correlates more strongly with workload than with performance and to integrate this finding in our thinking about usability, UX, and design. For example, the technology acceptance model (e.g., Venkatesh et al., 2003) states that effort expectancy, a construct that resembles workload, has a weaker effect on users' intention to use a system than performance expectancy has. Furthermore, the effect of effort expectancy tends to wear off as users gain experience with the system. That is, the relation between effort expectancy and intention to use is quite different from that between workload and preference even though effort expectancy and workload are similar in that both constructs measure the effort the user must expend to complete tasks with a system. The difference accentuates that expectations, such as effort expectancy, differ from experiences, such as experienced workload. This explanation suggests that when effort becomes salient as workload experienced during actual use, then its influence on preference increases. Future studies should investigate how sensitive preferences are to changes in workload and performance, for example distinguishing between pre-use, novice use, and experienced use. Existing studies have mostly investigated how preferences evolve from pre-use to novice-use situations (e.g., Lee and Koubek, 2010; Raita and Oulasvirta, 2011). This study adds that preference is more strongly correlated with workload, task time, and error rate for experienced than novice users. Follow-up studies could qualitatively analyze cases of dissociation to understand the psychological and contextual factors that drive preferences in these situations. Future studies should also investigate the extent to which users act on their preferences. How strong must a preference be before it leads to action?

The variation in workload, task time, and error rate collectively explains 57 % of the variation in preference, thereby leaving 43 % of the variation unexplained. To understand preference more fully, future studies must look beyond workload, task time, and error rate. The technology acceptance literature suggests that facilitating conditions, social influence, and perceived enjoyment explain some of the unexplained variation (Hornbæk and Hertzum, 2017; Venkatesh et al., 2003). Facilitating conditions are about the organizational and technical infrastructure in place to support the use of a system and may, thus, overlap somewhat with the workload users experience and the performance they achieve. In contrast, social influence and perceived enjoyment appear to be factors that are genuinely different from workload, task time, and error rate. While social influence indicates that preferences are co-determined by norms, peer groups, and other inter-personal issues, perceived enjoyment indicates that preferences are also co-determined by aesthetics, fun, and other non-utilitarian issues. It would be informative to learn how enjoyment and social norms influence preference, not just for hedonic and collaborative systems but also for interfaces for such mundane tasks as text entry.

5.4. Limitations

Four limitations should be remembered in interpreting the results of this study. First, the number of studies in this meta-analysis is limited by the inclusion criteria that eligible studies must report measures of preference, workload, task time, and error rate. The requirement that studies report all four measures was necessary to analyze how preference balances workload, task time, and error rate against one another. Additional studies report preference in combination with some, but not all, of the three other measures. Second, the information reported in the included studies restricted the choice of effect-size measure that could be employed in the meta-analysis. In particular, many of the studies did not report standard deviations. While the chosen effect-size measure was also used by Nielsen and Levy (1994), a measure that incorporated

information about the distribution of the data (i.e., standard deviations) would have been more robust. Third, workload is in this study self-reported using the TLX instrument. However, workload can also be measured in other ways, such as analytically, physiologically, and by means of a secondary task (Gawron, 2019). By not being self-reported, these other workload metrics measure it in ways that are shielded from how preference is measured. In the present study both preference and workload are self-reported. Future studies should validate the findings of this study with workload metrics that are not self-reported. Fourth, it is important to remember that this study is based on correlational data. While it appears most likely that preferences are the result of workload, task times, and error rates, this direction of influence remains an assumption. It cannot be ruled out that preferences influence behavior in ways that impact workload, task time, and error rate. The statements about the percentage of variation explained do not express causation; they are merely a more intuitive way of expressing correlations.

6. Conclusion

Low workload, task time, and error rate increase the probability that users prefer a system but do not guarantee it. Across the 144 studies in this meta-analysis, 60 % show a positive association between preference and all three of workload, task time, and error rate. However, in 2 % of the studies the preferred system comes with significantly higher workload than the nonpreferred system, in 8 % with significantly higher task time, and in 7 % with significantly higher error rate. With stronger preference, the difference in workload, task time, and error rate becomes larger. This correlation is strongest for workload; the variation in workload explains 54 % of the variation in preference. In contrast, the variation in task time and error rate explains 36 % and 19 %, respectively, of the variation in preference. That is, workload is a stronger predictor of preference than performance is. Workload correlates more strongly with preference for experienced than novice users and less strongly for safety-critical than non-safety-critical domains, suggesting that users accept higher workload when they are learning to use a system and when safety is at stake.

Researchers are invited to embrace the finding that preference correlates more strongly with workload than with performance, even for systems directed at utilitarian tasks, and to integrate this finding in our thinking about usability, UX, and design. Future studies should also investigate the relation between preference and workload in more detail and look for variables that explain additional variation in preference. For practitioners, the implications of this study are that it is risky to infer preference from performance, or vice versa, and that workload should be included in system evaluations.

Funding Sources

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Morten Hertzum: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Andre, A.D., Wickens, C.D., 1995. When users want what's not best for them. Ergon. Des. 3, 10–14. https://doi.org/10.1177/106480469500300403.

Backhaus, N., Trapp, A.K., Thüring, M., 2018. Skeuomorph versus flat design: user experience and age-related preferences. In: Proceedings of the DUXU2018 Conference on Design, User Experience, and Usability. Cham. Springer, pp. 527–542. https://doi.org/10.1007/978-3-319-91803-7_40.

Bailey, R.W., 1993. Performance vs. preference. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 37, 282–286. https://doi.org/10.1177/154193129303700406.

Bergman, O., Beyth-Marom, R., Nachmias, R., Gradovitch, N., Whittaker, S., 2008. Improved search engines and navigation preference in personal information management. ACM Trans. Inf. Syst. 26, 1–24. https://doi.org/10.1145/1402256.1402259.

Brooke, J., 1996. SUS: A "quick and dirty" usability scale, in: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, I.L. (Eds.), Usability Evaluation in Industry. Taylor & Francis, London, pp. 189–194.

Byers, J.C., Bittner, A.C., Hill, S.G., 1989. Traditional and raw task load index (TLX) correlations: are paired comparisons necessary? In: Mital, A. (Ed.), Advances in Industrial Ergonomics & Safety. Taylor & Francis, London, pp. 481–485.

Castillo, L., Permin, E., Fischer, J., Pyschny, N., 2023. A comparative study of digital assembly assistance systems. In: Proceedings of the CLF2023 Conference on Learning Factories. SSRN. Rochester, NY, pp. 1–7. https://doi.org/10.2139/ssrn.4469555.

Chen, Z., Malpani, A., Chalasani, P., Deguet, A., Vedula, S.S., Kazanzides, P., Taylor, R. H., 2016. Virtual fixture assistance for needle passing and knot tying. In: Proceedings of the IROS2016 Conference on Intelligent Robots and Systems. New York. IEEE, pp. 2343–2350. https://doi.org/10.1109/IROS.2016.7759365.

de Winter, J.C.F., 2014. Controversy in human factors constructs and the explosive use of the NASA-TLX: a measurement perspective. Cogn. Technol. Work 16, 289–297. https://doi.org/10.1007/s10111-014-0275-1.

Druckman, J.N., Lupia, A., 2000. Preference formation. Annu. Rev. Polit. Sci. 3, 1–24. https://doi.org/10.1146/annurev.polisci.3.1.1.

Fehige, C., Wessels, U. (Eds.), 1998. Preferences. De Gruyter, Berlin. https://doi.org/10.1515/9783110804294.

Frøkjær, E., Hertzum, M., Hornbæk, K., 2000. Measuring usability: are effectiveness, efficiency, and satisfaction really correlated?. In: Proceedings of the CHI2000 Conference on Human Factors in Computing Systems. New York. ACM, pp. 345–352. https://doi.org/10.1145/332040.332455.

Gawron, V.J., 2019. Human Performance, Workload, and Situational Awareness Measures Handbook, Third Edition. CRC Press, Boca Raton, FL. https://doi.org/10.1201/9780429019562.

Hart, S.G., 2006. NASA-task load index (NASA-TLX): 20 years later. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 50, 904–908. https://doi.org/10.1177/154193120605000909.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (task load index): results of empirical and theoretical research, in: Hancock, P.A., Meshkati, N. (Eds.), Human Mental Workload. North-Holland, Amsterdam, pp. 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9.

Hertzum, M., 2022. Associations among workload dimensions, performance, and situational characteristics: a meta-analytic review of the task load index. Behav. Inf. Technol. 41, 3506–3518. https://doi.org/10.1080/0144929X.2021.2000642.

Hertzum, M., 2021. Reference values and subscale patterns for the task load index (TLX): a meta-analytic review. Ergonomics 64, 869–878. https://doi.org/10.1080/00140139.2021.1876927.

Hertzum, M., Holmegaard, K.D., 2013. Perceived time as a measure of mental workload: effects of time constraints and task success. Int. J. Hum. Comput. Interact. 29, 26–39. https://doi.org/10.1080/10447318.2012.676538.

Hertzum, M., Hornbæk, K., 2010. How age affects pointing with mouse and touchpad: a comparison of young, adult, and elderly users. Int. J. Hum. Comput. Interact. 26, 703–734. https://doi.org/10.1080/10447318.2010.487198.

Hornbæk, K., 2006. Current practice in measuring usability: challenges to usability studies and research. Int. J. Hum. Comput. Stud. 64, 79–102. https://doi.org/10.1016/j.ijhcs.2005.06.002.

Hornbæk, K., Hertzum, M., 2017. Technology acceptance and user experience: a review of the experiential component in HCI. ACM Trans. Comput. Interact. 24, 33. https://doi.org/10.1145/3127358 article.

Hornbæk, K., Law, E.L.-C., 2007. Meta-analysis of correlations among usability measures. In: Proceedings of the CHI2007 Conference on Human Factors in Computing Systems. New York. ACM, pp. 617–626. https://doi.org/10.1145/1240624.1240722.

Hu, Y., Malthaner, R.A., 2007. The feasibility of three-dimensional displays of the thorax for preoperative planning in the surgical treatment of lung cancer. Eur. J. Cardio-Thoracic Surg. 31, 506–511. https://doi.org/10.1016/j.ejcts.2006.11.054.

ISO 9241-210, 2019. Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems. Int. Standard Org.

Jardina, J.R., Chaparro, B.S., 2012. Usability of e-readers for book navigation tasks. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 56, 1897–1901. https://doi.org/10.1177/1071181312561276.

Jeon, M., Davison, B.K., Nees, M.A., Wilson, J., Walker, B.N., 2009. Enhanced auditory menu cues improve dual task performance and are preferred with in-vehicle

technologies. In: Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications. New York. ACM, pp. 91–98. https://doi.org/10.1145/1620509.1620528.
Kim, H., Kwon, S., Heo, J., Lee, H., Chung, M.K., 2014. The effect of touch-key size on the usability of in-vehicle information systems and driving safety during simulated driving. Appl. Ergon. 45, 379–388. https://doi.org/10.1016/j.apergo.2013.05.006.
Kiss, F., Woźniak, P.W., Scheerer, F., Dominiak, J., Romanowski, A., Schmidt, A., 2019. Clairbuoyance: improving directional perception for swimmers. In: Proceedings of the CHI2019 Conference on Human Factors in Computing Systems. New York. ACM, pp. 1–12. https://doi.org/10.1145/3290605.3300467.
Kujala, T., 2013. Browsing the information highway while driving: three in-vehicle touch screen scrolling methods and driver distraction. Pers. Ubiquitous Comput. 17, 815–823. https://doi.org/10.1007/s00779-012-0517-2.
Kytö, M., Ens, B., Piumsomboon, T., Lee, G.A., Billinghurst, M., 2018. Pinpointing: precise head- and eye-based target selection for augmented reality. In: Proceedings of the CHI2018 Conference on Human Factors in Computing Systems. New York. ACM, pp. 1–14. https://doi.org/10.1145/3173574.3173655.
Lane, D.M., Napier, H.A., Peres, S.C., Sandor, A., 2005. Hidden costs of graphical user interfaces: failure to make the transition from menus and icon toolbars to keyboard shortcuts. Int. J. Hum. Comput. Interact. 18, 133–144. https://doi.org/10.1207/s15327590ijhc1802_1.
Lee, S., Koubek, R.J., 2010. Understanding user preferences based on usability and aesthetics before and after actual use. Interact. Comput. 22, 530–543. https://doi.org/10.1016/j.intcom.2010.05.002.
Lewis, J.R., Erdinc, O., 2017. User experience rating scales with 7, 11, or 101 points: Does it matter? J. Usability Stud. 12, 73–91.
Lindgaard, G., Fernandes, G., Dudek, C., Brown, J., 2006. Attention web designers: you have 50 milliseconds to make a good first impression! Behav. Inf. Technol. 25, 115–126. https://doi.org/10.1080/01449290500330448.
Littell, J.A., Corcoran, J., Pillai, V., 2008. Systematic reviews and meta-analysis. Oxford University Press, Oxford.
Liu, F., Zhu, Z., Chen, H., Li, X., 2020. Beauty in the eyes of its beholders: effects of design novelty on consumer preference. J. Retail. Consum. Serv. 53, 101969. https://doi.org/10.1016/j.jretconser.2019.101969.
Longo, L., 2018. Experienced mental workload, perception of usability, their interaction and impact on task performance. PLoS One 13, e0199661. https://doi.org/10.1371/journal.pone.0199661.
Mathis, F., Williamson, J.H., Vaniea, K., Khamis, M., 2021. Fast and secure authentication in virtual reality using coordinated 3D manipulation and pointing. ACM Trans. Comput. Interact. 28. https://doi.org/10.1145/3428121 article 6.
Matthes, J., 2007. Beyond accessibility? Toward an on-line and memory-based model of framing effects. Communications 32, 51–78. https://doi.org/10.1515/COMMUN.2007.003.
Mott, M.E., Wobbrock, J.O., 2014. Beating the bubble: using kinematic triggering in the bubble lens for acquiring small, dense targets. In: Proceedings of the CHI2014 Conference on Human Factors in Computing Systems. New York. ACM, pp. 733–742. https://doi.org/10.1145/2556288.2557410.
Nielsen, J., Levy, J., 1994. Measuring usability: preference vs. performance. Commun. ACM 37, 66–75. https://doi.org/10.1145/175276.175282.
Nygren, T.E., 1991. Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. Hum. Factors 33, 17–33. https://doi.org/10.1177/001872089103300102.
Özacar, K., Hincapié-Ramos, J.D., Takashima, K., Kitamura, Y., 2016. 3D selection techniques for mobile augmented reality head-mounted displays. Interact. Comput. 29, 579–591. https://doi.org/10.1093/iwc/iww035.
Prilla, M., Mantel, A.M., 2021. Analysing a UI's impact on the usability of hands-free interaction on smart glasses. In: Proceedings of the ISMAR2021 International Symposium on Mixed and Augmented Reality. New York. IEEE, pp. 421–426. https://doi.org/10.1109/ISMAR-Adjunct54149.2021.00095.
Quinn, P., Cockburn, A., 2020. Loss aversion and preferences in interaction. Hum.–Comput. Interact. 35, 143–190. https://doi.org/10.1080/07370024.2018.1433040.
Raita, E., Oulasvirta, A., 2011. Too good to be bad: favorable product expectations boost subjective usability ratings. Interact. Comput. 23, 363–371. https://doi.org/10.1016/j.intcom.2011.04.002.
Ranasinghe, C., Schiestel, N., Kray, C., 2019. Visualising location uncertainty to support navigation under degraded GPS signals: a comparison study. In: Proceedings of the MobileHCI2019 Conference on Human-Computer Interaction with Mobile Devices and Services. New York. ACM, pp. 1–11. https://doi.org/10.1145/3338286.3340128.
Reid, G.B., Nygren, T.E., 1988. The subjective workload assessment technique: a scaling procedure for measuring mental workload, in: Hancock, P.A., Meshkati, N. (Eds.), Human Mental Workload. North-Holland, Amsterdam, pp. 185–218. https://doi.org/10.1016/S0166-4115(08)62387-0.
Sauro, J., Lewis, J.R., 2009. Correlations among prototypical usability metrics: evidence for the construct of usability. In: Proceedings of the CHI2009 Conference on Human Factors in Computing Systems. ACM, New York, pp. 1609–1618. https://doi.org/10.1145/1518701.1518947.
Schenkman, B.N., Jönsson, F.U., 2000. Aesthetics and preferences of web pages. Behav. Inf. Technol. 19, 367–377. https://doi.org/10.1080/014492900750000063.
Schramm, R.C., Sasalovici, M., Hildebrand, A., Schwanecke, U., 2023. Assessing augmented reality selection techniques for passengers in moving vehicles: a real-world user study. In: Proceedings of the 15th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. New York. ACM, pp. 22–31. https://doi.org/10.1145/3580585.3607152.
Thomas, R.C., 1998. Long term human-computer interaction: An exploratory perspective. Springer, London. https://doi.org/10.1007/978-1-4471-1548-9.
Turner, C.J., Chaparro, B.S., He, J., 2021. Typing on a smartwatch while mobile: a comparison of input methods. Hum. Factors 63, 974–986. https://doi.org/10.1177/0018720819891291.
Venkatesh, V., Morris, M.G., Davis, G.B., Davis, F.D., 2003. User acceptance of information technology: Toward a unified view. MIS Q 27, 425–478. https://doi.org/10.2307/30036540.