Anchoring Effects and Troublesome Asymmetric Transfer in Subjective Ratings

Andy Cockburn
University of Canterbury
Christchurch, New Zealand
andy@cosc.canterbury.ac.nz

Carl Gutwin
University of Saskatchewan
Saskatoon, Canada
gutwin@cs.usask.ca

ABSTRACT

Within-subjects experiments are prone to asymmetric transfer, which confounds results interpretation. While HCI researchers routinely test asymmetric transfer in objective data, doing so for subjective data is rare. Yet literature suggests that *anchoring effects* should make subjective measures particularly susceptible to asymmetric transfer. We report on four analyses of NASA-TLX data from four previously published HCI papers, with four main findings. First, asymmetric transfer is common, occurring in 42% of tests analysed. Second, the data conforms to predictions of anchoring effects. Third, the magnitude of the anchor's effect correlates with the magnitude of the difference between the interface ratings – that is, the anchor's 'pull' correlates with the anchoring stimulus. Fourth, several of the previously published findings are changed when data are reanalysed using between-subjects treatment. We urge caution when analysing within-subjects subjective measures and recommend that researchers test for and report the occurrence of asymmetric transfer.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; **Human computer interaction (HCI)**;

KEYWORDS

Anchoring effects, asymmetric transfer, order effects, subjective measures, NASA-TLX.

ACM Reference Format:
Andy Cockburn and Carl Gutwin. 2019. Anchoring Effects and Troublesome Asymmetric Transfer in Subjective Ratings. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland Uk.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300592

CHI 2019, May 4–9, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300592

INTRODUCTION

Controlled experiments play a pivotal role in the development of foundational understanding in Human-Computer Interaction. Subjects are assigned to experimental conditions, with objective and subjective data taken to help researchers understand various aspects of technology and human reactions to it. Objective data normally characterises aspects of performance, such as task time or error rates; whereas subjective data provides insights into human experiential factors, such as perceived workload, frustration, preferences, and so on. To assist researchers in gathering subjective data, and to facilitate comparison across studies, various survey techniques have been developed, such as the Questionnaire for User Interface Satisfaction (QUIS) [6] and the NASA Task Load Index (NASA-TLX) [15]. These survey methods typically require users to select a response from a small set of ordered values on a semantic differential scale or Likert item, with responses translated into corresponding numerical values (e.g., 1 to 5). The existence of large sets of HCI studies using similar scales facilitates meta-analyses that would otherwise be impossible. For example, through a meta-analysis of 127 different user interface evaluations, Nielsen and Levy [27] determined that the median response for interface satisfaction was 3.6 out of 5. In short, quantitative analysis of subjective measures plays an important role in HCI research.

HCI experiments are often designed to use within-subjects treatment for interface factors, meaning that all experimental participants complete tasks using all of the interfaces evaluated (as opposed to between-subjects treatment in which each participant uses exactly one interface). Two reasons for preferring within-subjects treatment are that, first, it provides some control for individual variability (a user who is particularly slow with one interface is likely to be comparatively slow with other interfaces) and second, it facilitates participant economy (one participant provides data for n interfaces, whereas between-subjects assignment requires n participants to gather the same quantity of data). The primary drawback of within-subjects treatment involves *order effects* – there are risks that unintended factors such as learning or

fatigue effects may influence measures across conditions. For example, when an experimental participant completes tasks using interface *A* and then *B*, their faster performance with *B* may be explained by *B* providing better support than *A*; or perhaps the improvement is due to the participant being more familiar with the experimental tasks (a learning effect).

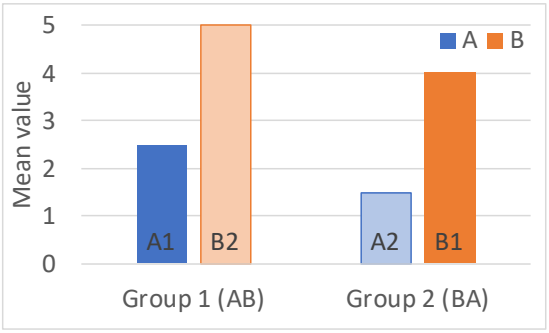
To mitigate confounding order effects, researchers use order *counterbalancing* – for example, half of the participants might complete condition *A* before *B*, with the other half completing *B* before *A*. When experiments involve more than two conditions, more complex counterbalancing measures can be applied, such as using a Latin square to assure that no condition prevalently precedes another.

However, counterbalancing does not eliminate risks of order effects. In particular, there are risks of *asymmetric transfer* [31] in which the order of experience differently influences the transfer effect between conditions. Figure 1a shows a hypothetical example of asymmetric transfer (the meaning of the scale is unimportant). In Group 1 participants completed tasks in the order A_1 then B_2 (meaning *A* first, *B* second), with condition A_1 serving as good preparation for condition B_2 (mean ratings of 2.5 and 5 respectively). However in Group 2 (order B_1 then A_2), B_1 with a mean of 4 served as poor preparation for A_2 resulting in it receiving a mean value of 1.5.

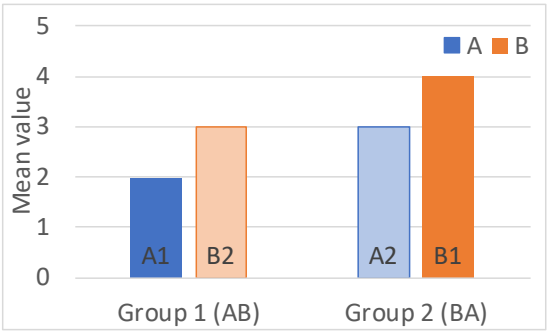
Asymmetric transfer can be detected through a group-based analysis of the dependent measure – for example, comparing the mean value for the A_1B_2 group with that of the B_1A_2 group. When asymmetric transfer occurs, appropriate results interpretation is fraught, and researchers will typically discard all data except that from the first experienced condition, using between-subjects analysis.

The need to be cautious about asymmetric transfer in within-subjects experiments is well known in HCI, featuring in texts on research methodology (e.g., [23]). However, while we are aware of many studies that have inspected *objective* data for the presence of asymmetric transfer, we are unaware of studies in HCI that have done so for their subjective data.

Importantly, psychology research on *anchoring effects* suggests that subjective responses are particularly susceptible to asymmetric transfer. Anchoring effects cause people to import earlier numerical experiences into their subsequent numerical assessments – the earlier number becomes an anchor that influences the latter number. Tversky and Kahneman [38] demonstrated the effect in an experiment in which participants first watched a roulette wheel that was rigged to have the ball land on either 10 or 65; and subsequently participants guessed the proportion of African countries in the United Nations. The mean guesses were 25% following 10 and 45% following 65. The anchor values 10 and 65 'pulled' responses down and up, respectively.



(a) A_1 **raises** B_2; B_1 **lowers** A_2.



(b) A_1 **lowers** B_2; B_1 **raises** A_2.

Figure 1. Two hypothetical forms of asymmetric transfer. Top, A_1 before B_2 raises B_2's value, but B_1 before A_2 lowers A_2's value. Bottom, consistent with anchoring effects, in Group 1, A_1's low value drags down B_2's value, but in Group 2, B_1's high value before A_2 drags up A_2's value.

Figure 1b characterises how anchoring effects might influence users' subjective assessments of interface experiences. In the figure, the dark bars (A_1 and B_1) show the conditions completed first, representing 'truthful' subjective values that are unaffected by transfer. In Group 1, the low first assessment of A_1 results in an artificially low response for B_2 (the low value anchors the second assessment); and in Group 2, the high initial assessment of B_1 provides an anchor that lifts the subsequent assessment of A_2. Importantly, anchoring effects should result in asymmetric transfer (indicated in this example by a difference in means between Groups 1 and 2), which confounds results interpretation. Furthermore, anchoring effects should diminish the truthful differences between interfaces *A* and *B*, potentially leading to Type II errors.

This paper describes four analyses in which NASA-TLX data from four previously published within-subjects HCI experiments were reinspected. The specific contributions of the paper are as follows: 1. a review-based appraisal of the likelihood that anchoring effects will cause asymmetric transfer in user interface subjective ratings; 2. confirmation that asymmetric transfer occurred in many of the previously

published tests; 3. confirmation that the data conforms to anchoring effects; 4. results showing that the strength of the anchoring outcome correlates with the magnitude of the anchor's stimulus; and 5. evidence that several of the previously published findings would change if reanalysed in a between-subjects manner (discarding data from all but the first condition experienced). The important implications for future analysis of subjective measures are discussed.

BACKGROUND

Three areas of background work are briefly reviewed, concerning detection of asymmetric transfer, psychology foundations on anchoring and related effects, and prior studies in HCI examining these issues.

Detecting Asymmetric Transfer

Asymmetric transfer occurs when there is a differential transfer across experimental conditions dependent on order of exposure [31]. In user interface evaluation this might occur if interface A provided good preparation for B, while B provided poor preparation for A. Asymmetric transfer is detected when there is a statistically significant difference in the dependent measure between *group* ordering (i.e., the mean for the A_1B_2 group differs from that of the B_1A_2 group). The charts shown in Figure 1 both show this type of effect.

The occurrence of asymmetric transfer is sometimes inspected by investigating the interaction between factors $order \in \{first, second\}$ and *condition* (such as interface A and B). However this method is logically equivalent to testing the main effect of group. For example, the data from Figure 1a can be redrawn (Figure 2) to show an interaction of *order*×*interface*, with A values decreasing from A_1 to A_2 while B values increase. But conducting the analysis in this way is problematic due to incomplete cells in the experimental design. This occurs because participants complete both orders ($first$ and *second*) and both interfaces (A and B), suggesting that both factors are within-subjects, but the factors are not crossed, meaning that for each participant only one of A or B is completed first.

The general problems of asymmetric transfer [31] are well documented in the HCI literature for objective measures (e.g., [22]), but not for subjective measures. Furthermore, the problems of cross-contamination to within-subjects subjective responses have been discussed in relation to the testing of utility theories in behavioural economics [20], and this knowledge has influenced experimental designs within HCI (e.g. [32]). However, we are unaware of prior HCI research that has analysed or demonstrated asymmetric transfer for subjective measures.

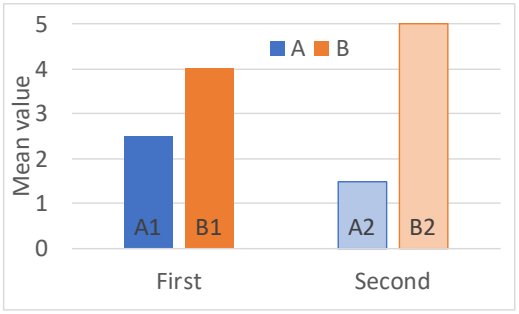


Figure 2. Asymmetric transfer evident in an *Order*×*Interface* interaction, using identical data to that shown in Figure 1a: values for A decrease from First to Second, while values for B increase.

Foundations on Anchoring and Related Effects

The findings of reliable anchoring effects have been confirmed many times since Tversky and Kahneman's [38] famous roulette wheel study. While many studies have focussed on the corrupting influence of an anchor value on quantitative subjective judgements, others have generalised these findings to subjective valuation and preference. For example, Ariely [2] conducted a study that asked subjects whether they would be willing to buy items such as wine or books for a value equal to the last two digits of their social security number, and they were subsequently asked to state their maximum willingness-to-pay (WTP) value. Higher social security numbers resulted in higher WTP values – for example, subjects whose social security number was in the top quintile were willing to pay an average of $56 for a cordless computer keyboard, compared to only $16 for those whose numbers were in the bottom quintile. In another of his studies, Ariely examined the price at which subjects would be willing to accept (WTA) re-listening to a brief painful noise played over headphones. Higher initial offers (higher anchors) resulted in higher payments required before reaching a WTA threshold. Anchoring effects have also been shown to influence valuation in art auctions [5], general purchase and selling decisions [37], and even criminal sentences in judicial decisions [12]. A full review of literature on anchoring effects is beyond the scope of this paper; interested readers are directed to Furnham and Boo's relatively recent review [13].

In summarising anchoring effects, Ariely stated 'an initial choice will exert a normatively inappropriate influence over subsequent choices and values' [2, p78]. If true for HCI experiments involving within-subjects use of subjective rating scales, this is a clear area for concern.

Other psychological effects could also influence users' subjective responses when engaged in a series of assessments. In particular, Kahneman and Tversky's *prospect theory* [19]

states that people make judgements with respect to a reference point (e.g., expecting to be paid $50 for some work). If an outcome is less than the reference point (e.g., being paid $40), then it represents an objective loss, and if greater (e.g., being paid $60), it is an objective gain. A key transformation in prospect theory is that the magnitude of the subjective value associated with losses is greater than the subjective value associated with gains – being paid $40 feels worse than being paid $60 feels good. The general sentiment is that 'bad is stronger than good' (see Baumeister et al. [4] for a review).

Importantly for our discussion, reference points can shift [21]. Consequently, if the 'good' experience of one interface caused the reference point to rise (raising expectations), then the failure of the next interface to meet the new reference point could result in an amplified negative subjective value derived from the loss. This hypothetical scenario is characterised in Figure 1a – in Group 1, the initial low value of A_1 lowers the reference point, resulting in a higher than normal assessment of B_2; and in Group 2, the high assessment of B_1 raises the reference point, resulting in a lower than normal assessment of A_2 (lowering the overall mean for A). Note that the hypothetical influence of a shifting reference point pulls subsequent assessment values in the opposite direction to that of the hypothetical influence of anchoring. Under a theory of shifting reference points, the difference between the two second interfaces grows ($|A_2 - B_2| > |A_1 - B_1|$, Figure 1a), potentially promoting Type I errors of false identification of a significant difference.

In contrast, under a theory of anchoring, the equivalent difference diminishes ($|A_2 - B_2| < |A_1 - B_1|$, Figure 1b). While the direction of the difference between assessments of A and B should not change under anchoring, any reduction in the magnitude of the difference between scores of A and B could potentially cause Type II errors (failures to identify a significant difference) for two reasons. First, the widely used Wilcoxon sign-rank test incorporates data on the direction and magnitude of difference, so reductions in the magnitude of difference are likely to reduce test sensitivity. Second, survey response scales are typically coarse in granularity (typically five or seven items), so a reduction in the magnitude of difference may increase the proportion of score ties, again reducing the likelihood that a significant difference will be identified. Furthermore, papers frequently report mean and other values for scale responses to indicate differences between conditions, so it would be useful to know if systematic effects underestimate these differences.

Primacy effects [3, 11, 30] could also influence users' responses to a series of survey questions. Primacy effects show that the probability of recalling items in a series of cued items (such as words) follows a U-shape across serial position, with initial and terminating items recalled best. This effect has been generalised to encompass a well-validated *priming* tendency for information that is acquired early to have a disproportionally strong influence on memories, decisions, and judgements (e.g., [26]). Close connections are suspected between the underlying mechanisms of primacy, priming, and anchoring, but remain under debate (e.g., [25, 26, 38]).

Finally, *straight-lining* effects have been shown to occur [17], particularly in long surveys, with respondents providing identical responses to questions regardless of their content.

Anchoring and Related Effects in HCI

Anchoring and related psychological effects have been studied in various areas of HCI, with work on recommender systems making the most direct and frequent reference to the effects.

Cosley *et al.* [10] examined how recommender systems influence users' subjective ratings. Results from their laboratory study of a movie-rating interface showed that users' ratings were significantly lower or higher than normal when the interface showed a predicted rating that was respectively lower or higher. These results conform to an anchoring effect, although the paper makes no mention of anchoring. Adomavicius [1] later noted that Cosley *et al.*'s results could be explained by anchoring effects, and Zhang *et al.* concluded that 'the rating provided by a recommender system serves as an anchor for the consumer's constructed preference' [42, p375]. Similar effects have also been observed in information retrieval [36] (user's relevance assessment of documents is influenced by the quality of the last document judged) and in information visualization [39] (the interpretation of a visualization can be influenced by the previous visualization).

These studies examined effects influencing the user's assessment of information presented through the interface. However, previous HCI studies have also examined the role of related effects on assessments of the interface itself. For example, Hartmann *et al.* [16] examined a variety of influences on users' assessment of interface aesthetics, finding that ratings tended to spill over between assessment categories in a form of halo effect [28] that might also be influenced by anchoring: 'attribution of good quality on one attribute positively influenced judgment on another, even in the face of objective evidence to the contrary' [16, p15]. Similarly, studies by Raita and Oulasvirta [33] showed that a positive priming stimulus increased interface usability ratings, while negative primes reduced them. Michalco *et al.* review several studies showing related effects, placing their own studies within the framework of expectation disconfirmation [29]. Although an interesting framework, expectation disconfirmation has been criticised because its underlying theories can be used to predict all possible experimental outcomes while prohibiting none [41].

ANALYSIS ONE: ASYMMETRIC TRANSFER

To examine whether asymmetric transfer occurred in past studies, we required data sets in which two interface conditions were examined using within-subjects treatment, and with interface order recorded. We selected studies using the NASA Task Load Index (NASA-TLX) [15] for three reasons: it is among the most widely used ratings methods for subjective assessment, it provides six different data-points (mental, physical and temporal workload, success/performance, overall effort, and frustration), and it has relatively consistent means for presentation in the form of numerical scales, typically in the range 1-5, 1-7 or 1-9.

We limited our analyses to studies that investigated exactly two interface conditions. We did so because analyses with three or more conditions would likely produce tiny samples for the between-subjects analysis that is necessary when considering group ordering effects. For example, a study with 18 participants across three interface conditions has six different potential orderings, which provides only three participants' data in each order.

Data was retrieved from four of our own previously published studies – [24], with $n = 36$, and responses gathered using paper survey sheets on a 5-point scale; [8], $n = 28$, paper 7-point scale; [35], $n = 10$, paper 5-point scale; and [14], $n = 16$, online 10-point scale. The source data file and R script are available on the ACM Digital Library as supplementary material. The four studies yield 48 separate subjective measurements (4 studies × 6 NASA-TLX measures × 2 interface conditions), and the same number of group ordering data points ($4 \times 6 \times 2$ groups (A_1B_2 and B_1A_2)). We normalised data from all studies to real values in the range 1-5.

We used the ARTool R package to apply the Aligned Rank Transform [18, 34] to NASA-TLX measures. Having done so, we used a mixed two factor ANOVA to analyse the effect of *Group* $\in \{A_1B_2, B_1A_2\}$ (between-subjects) and *Interface* $\in \{A, B\}$ (within-subjects) on the NASA-TLX measures. To reiterate, the subscript in the coding A_1 means that interface A was conducted first. As stated above, asymmetric transfer was identified though a significant effect of *Group*.

Results: Confirming Asymmetric Transfer

Figure 3 summarises the mean subjective ratings for Group 1 (A_1 then B_2, left in each plot) and Group 2 (B_1 then A_2, right in each plot) in the six NASA-TLX measures (columns) across the four studies (rows). The p value for the main effect of *Group* in each of the 24 separate ANOVAs is shown within the plots (p_g), in green bold text where significant, indicating asymmetric transfer.

In total, 10 of the 24 tests (42%) indicated significant asymmetric transfer at $\alpha = .05$ – the 42% value represents the proportion of tests that would have indicated significant asymmetric transfer if the original researchers had tested for its occurrence. All of the four studies showed asymmetric transfer in at least two tests. Of course, with 24 tests, at $\alpha = .05$ we should expect at least one false positive (Type I error) among the data – Bonferroni-Holm correction for a hindsight-based analysis indicates a 33% detection rate. This rate of asymmetric transfer should not occur without some underlying cause. Frustration measures showed significant asymmetric transfer in all four studies (none for mental effort), but the number of studies examined is too small to indicate whether differences exist across NASA-TLX categories.

ANALYSIS TWO: EVIDENCE OF ANCHORING

Anchoring effects should cause a particular form of asymmetric transfer in which a rating that follows a low initial assessment is pulled down by the low anchor, and a rating that follows a high initial assessment is pulled up. To determine whether the data was consistent with anchoring effects, we made two initial computations, as illustrated in Figure 4.

First, we determined which condition (A_1 or B_1) provided the 'low' anchor and which provided the 'high' anchor. We did so by comparing $\bar{A}_1$ with $\bar{B}_1$ (the mean values when conducted first, and therefore uninfluenced by transfer effects).

Second, we calculated the magnitude and direction of the resultant 'pull' of the anchor on the second assessment. This involved calculating the difference between the second and first mean rating for the interfaces: for example, $\Delta\bar{A} = \bar{A}_2 - \bar{A}_1$, as shown in Figure 4.

If anchoring effects apply, then the Δ value for the interface following the 'high' interface should be positive, and the Δ value should be negative when following the 'low' interface.

We conducted a paired t-test to compare the influence of 'low' and 'high' initial assessments on the subsequent assessment. In this test the dependent measure was the appropriate Δ value for the second interface: for example, in Figure 4, $\Delta\bar{B}$ is the dependent measure for the 'low' condition (because interface B followed the low value of A_1), and $\Delta\bar{A}$ is the dependent measure for the 'high' condition (because interface A followed the high value of B_1). Each of the 24 different NASA-TLX outcomes (four studies each with six workload measures) gives two values for the paired analysis: the Δ value following the 'low' assessment and the Δ value following the 'high' assessment.

Results: Support for Anchoring Effects

Results are summarised in Figure 5. The mean Δ value following the *low* interface was -0.34 (s.d. 0.62, 95% CI [-0.61, -0.07]), compared to a mean value of +0.53 (s.d. 0.58, 95% CI [0.28,

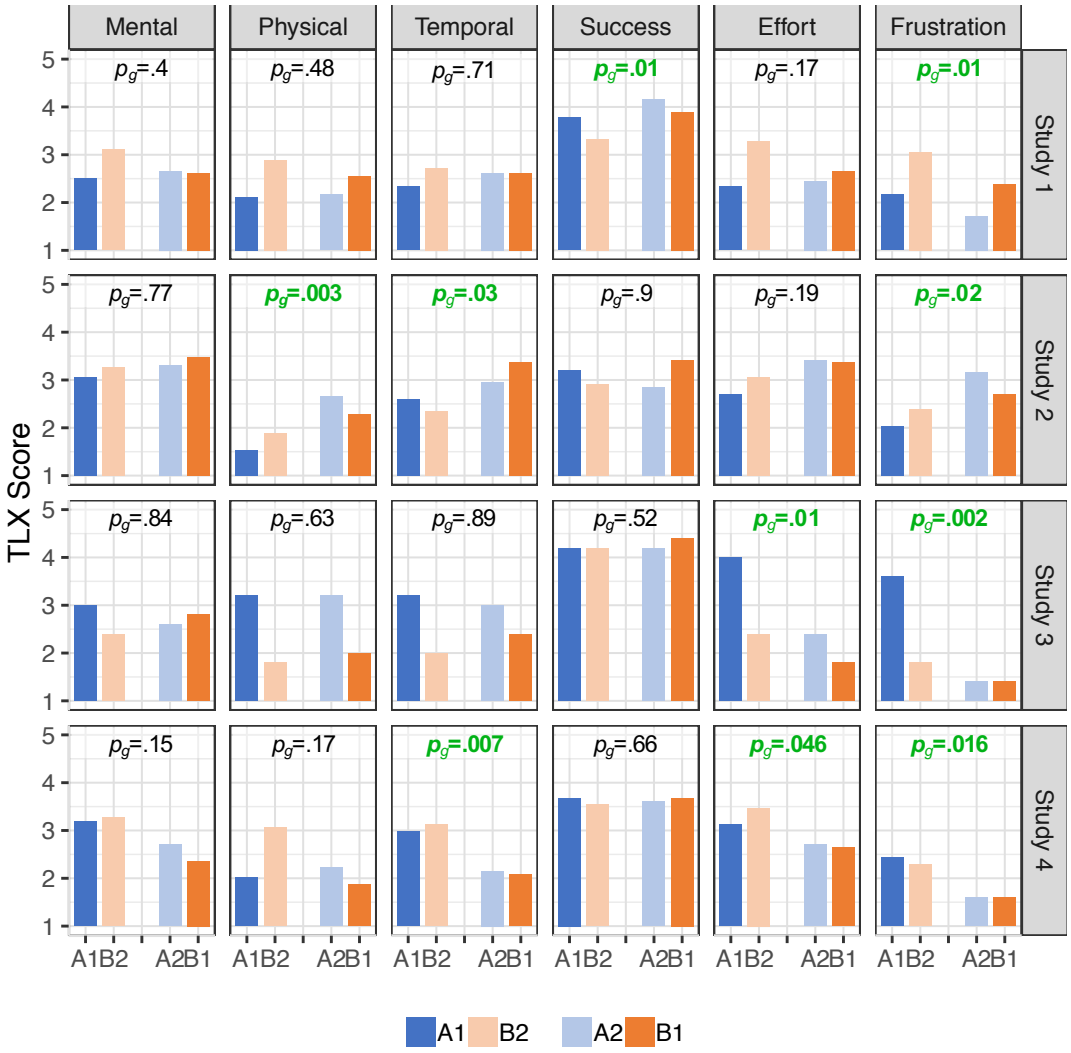


Figure 3. Mean values for each of the NASA-TLX measures in the four studies, together with p values for the main effect of *Group*. Green and bold values of p_g indicate significant asymmetric transfer at $\alpha = .05$.

0.78]) following the *high* interface: $T_{22} = 4.47$, $p < .0005$[1]. Note that the 95% confidence intervals following the *low* anchor only include negative values; and only positive values following the *high* anchor.

These results are consistent with anchoring effects – when users provide an initial low assessment, their subsequent assessment of the second interface is pulled down by the anchor value; and when they provide a high assessment, their subsequent assessment of the second interface is pulled up by the anchor value.

[1]The degrees of freedom are 22 rather than 23 because one test was excluded due to $\bar{A}_1 == \bar{B}_1$, preventing classification of either interface as *low*.

ANALYSIS THREE: ANCHOR STIMULUS & PULL

The analyses above indicate that asymmetric transfer is worryingly common in subjective ratings (42% in our sample) and that the nature of asymmetric transfer is consistent with anchoring effects. However, we were concerned that the rate of asymmetric transfer may be even higher when there are meaningful differences between the interfaces. For example, if two interfaces are truthfully similar to one another in their mental effort (say), then there is little stimulus for a resultant anchoring effect. By analogy, when assessing the value of a house with a nominal value of $1,000,000, anchoring cues of $300,000 and $3,000,000 provide a strong stimulus for different value assessments due to anchoring effects, but cues of

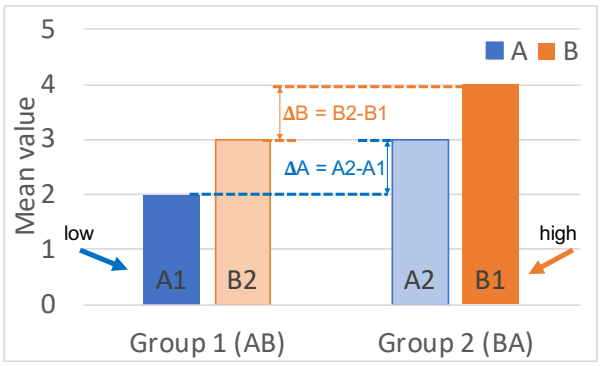


Figure 4. Illustrating the computations conducted to analyse evidence of anchoring effects – the 'low'/'high' anchor is the lower/higher of A_1 and B_1. Δ values represent the resultant 'pull' of the anchor – the difference between the second assessment value (potentially influenced by the anchor) and the value when assessed first.

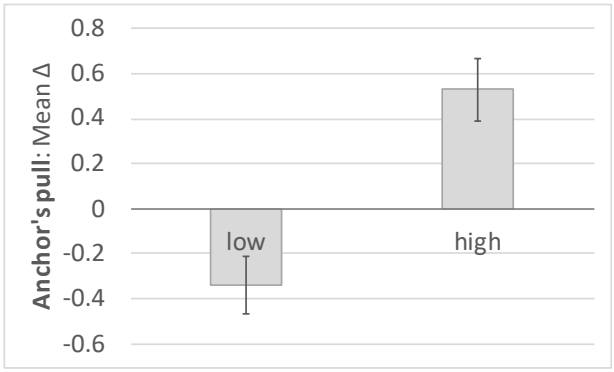


Figure 5. Mean Δ values (e.g., $\bar{A}_2 - \bar{A}_1$), representing the anchor value's pull on the assessment that follows the *low* (left) and *high* (right) first assessment. Error bars ± 1 s.e.m. Results conform to an anchoring effect, as characterised in Figure 4.

$300,001 and $300,002 provide a weak stimulus for different assessments.

We therefore used correlation analysis to investigate the relationship between the strength of anchoring stimulus and the upward or downward 'pull' of the anchor. Stimulus strength was calculated as the difference in mean rating values for the interfaces when completed first (e.g., $\bar{A}_1 - \bar{B}_1$), which was correlated with the Δ value for the interface that followed (the resultant 'pull').

Results: Stimulus Correlates with its 'Pull'

Figure 6 summarises the results, revealing a positive correlation between the strength of anchoring stimuli and the resultant 'pull' of the anchor value (Pearson's $r = .67$, classified as a 'large effect' according to Cohen's conventions [9]).

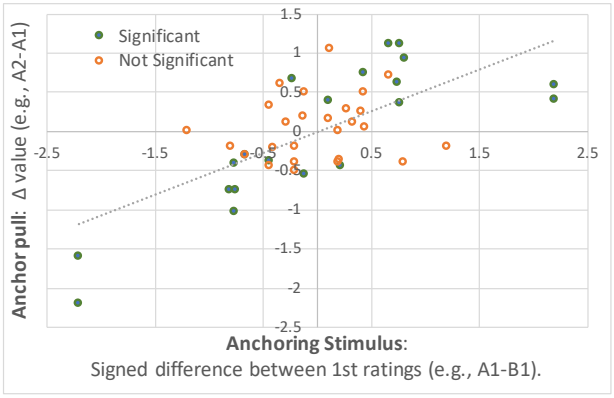


Figure 6. Correlation analysis of the mean strength of anchoring stimulus (the difference between A_1 and B_1 ratings, with *lower* interfaces left and *higher* interfaces right) and the resultant 'pull' of the anchor (Δ values for the interfaces). Data points corresponding to significant asymmetric transfer are green and filled.

In the figure, the filled green dots indicate conditions associated with significant asymmetric transfer (summarised in Figure 3); unfilled orange dots indicate no significant asymmetric transfer. If the correlation analysis is conducted incorporating only the data that showed significant asymmetric transfer, then Pearson $r = .82$; and the correlation is near absent ($r = .15$) when using only the data that showed no significant effect.

In other words, the strength of the anchoring stimuli correlates positively with the upwards or downwards pull on the subsequent rating for the equivalent measure in the next interface.

ANALYSIS FOUR: DO PRIOR RESULTS CHANGE?

The main concern about asymmetric transfer is that it confounds results interpretation and promotes false conclusions. We therefore conducted a final set of analyses to inspect whether the findings from the NASA-TLX results in previously published papers would change when accounting for asymmetric transfer.

We recalculated the original findings using within-subjects Wilcoxon signed ranked tests. We also computed new statistical measures using between-subjects Mann-Whitney tests of *only* the condition conducted first (i.e., A_1 or B_1, discarding data from A_2 and B_2), which removes the potential impact of asymmetric transfer.

Results: Many Previous Results Change

Table 1 summarises the findings for the six NASA-TLX measures across each of the four studies. Columns to the left show the originally published within-subjects analyses; the between-subjects reanalysis is to their right. Both of these

analyses show mean values for the two interfaces, the difference between these values (Δ_w and Δ_b for within- and between-subjects comparisons), a comparison of whether the mean for A is greater, less, or similar to B (within 0.1), and the resultant statistical p value. The 'A B Gap Change' columns indicate the difference between Δ values in the between- and within-subjects analyses – a negative value indicates that the difference between interfaces was smaller in the between-subjects difference than in the within-subjects analysis ('closing' the gap, which is the opposite of the predicted effect of anchoring); a positive value indicates that the between-subjects analysis shows a larger difference between interfaces than the within-subjects analysis (consistent with anchoring). The final column on the right indicates any change between the within- and between-subjects analyses.

We briefly highlight four different forms of change that occurred.

First, there are eleven cases of *gap widening*, where the difference between $\bar{A}_1$ and $\bar{B}_1$ in the new analysis is larger than the originally reported difference between $\bar{A}$ and $\bar{B}$. This effect occurs in Studies 2, 3, and 4. For example, the original values for Frustration in Study 3 showed a magnitude of difference of 0.9, which widens to 2.2 in the between-subjects analysis. This gap widening effect is consistent with anchoring – anchoring should diminish the truthful differences between conditions (see Figure 1b) because an initial high assessment pulls up the second assessment, and an initial low assessment pulls down the second assessment, falsely closing the gap.

Second, in four cases, the gap widening effect contributed to previously non-significant results becoming significant (Physical workload in Study 2, Effort and Frustration in Study 3, and Temporal workload in Study 4). In the cases of Physical Workload in Study 2 and Temporal Workload in Study 4 the meaning of the data is substantially altered, with the previous suggestion of equivalence between conditions being replaced with a finding that one interface is more demanding than the other. Three other analyses also suggest that this effect is occurring, with the between-subjects analyses approaching significance when the previously reported means were similar (Temporal workload and Frustration in Study 2, and Frustration in Study 4).

Third, there are five cases of *gap closing*, where the between-subjects analysis results in a smaller difference between $\bar{A}_1$ and $\bar{B}_1$ than originally shown between $\bar{A}$ and $\bar{B}$. This is evident in Study 1's measures for Mental effort, Success, Effort and Frustration, and in Study 4's measure of Physical effort. While it may appear that these data points oppose a theory of anchoring, the earlier analysis of the strength of the anchoring stimuli suggests that this is not the case. In all but one case where gap closing occurs, the magnitude of the difference between the 'truthful' $\bar{A}_1$ and $\bar{B}_1$ values is small (0.1 for Mental effort and Success in Study 1, and also for Physical effort in Study 4; 0.2 for Frustration in Study 1). These small differences suggest that anchoring effects should have had little influence because the initial assessments of both interfaces were similar.

Finally, there are four cases in Study 1 and one in Study 3 where originally significant findings become insignificant in the between-subjects analysis. One factor contributing to this is the loss of statistical power that comes from the loss of half of the sample due to between-subjects treatment, and the loss of control for individual variability that is inherent in within-subjects treatment. The most puzzling data point in this analysis, however, is Success in Study 1, where a marginal effect is shown with $\bar{A} > \bar{B}$ in the within-subjects analysis, despite similar values for $\bar{A}_1$ and $\bar{B}_1$. We currently have no explanation for why this analysis showed significance in the within-subjects treatment, suspecting that noise in random sampling is at play (a Type I error).

DISCUSSION

To summarise the main findings, first, reanalysis of four previously published studies showed a high rate of asymmetric transfer in NASA-TLX measures (42% at $\alpha = .05$). Second, the nature of the asymmetric transfer was analysed, showing that it conformed to anchoring effects. This analysis involved comparing the subjective ratings for the two interfaces when conducted first, treating one as providing a 'low' anchor and the other as a 'high' anchor, and then inspecting whether the second assessment rating for each interface was pulled up or down with respect to its first assessment value. Results showed that the 'low' anchor pulled subsequent values down, while the 'high' anchor pulled subsequent values up. Third, correlation analysis also supported anchoring effects, indicating that stronger anchoring stimuli (i.e., bigger differences between the subjective ratings of interfaces when assessed first) induced larger 'pulls' upwards or downwards. Finally, new between-subjects analysis (the standard practice when asymmetric transfer occurs) indicated that the basic meaning of several previously published findings would have changed if asymmetric transfer had been originally inspected.

The following discussion generalises the findings, offers suggestions for future researchers, and raises cautions for the research community.

Why and When Anchoring Occurs

The results show strong evidence for prevalent asymmetric transfer in subjective measures, with good indications that the effects are attributable to anchoring effects. However, further work is needed to better understand the nature and extent of these effects.

Table 1. Summary of published within-subjects analyses and new between-subjects analyses, including means for A and B, Δ between these values, order of values for A and B, and p values. The 'A B Gap Change' columns indicate the difference between Δ_b and Δ_w, with the 'Type' column showing whether the between-subjects analysis shows a smaller difference between interfaces than the within-subjects analysis ('closing' the gap, opposing anchoring) or 'widening' the gap (consistent with anchoring). Value differences of less than $|0.1|$ are interpreted as being similar ($\approx$). A tick in the 'Asym?' column shows that significant asymmetric transfer was indicated. The 'Meaning Change' column summarises any change of meaning between the published and new analyses.

Study	Published (within-subjects)					Reanalysed (between-subjects)					A B Gap Change		Asym?	Meaning Change
	$\bar{A}$	$\bar{B}$	Δ_w	Order	p	$\bar{A_1}$	$\bar{B_1}$	Δ_b	Order	p	$\lvert\Delta_b\rvert - \lvert\Delta_w\rvert$	Type		within → between
1														
Ment.	2.6	2.9	-0.3	$A < B$	.23	2.5	2.6	-0.1	$A \approx B$	.84	-0.2	Closing		$A < B \rightarrow A \approx B$
Phys.	2.1	2.7	-0.6	$A < B$	**.02**	2.1	2.6	-0.5	$A < B$	.18	-0.1	$\approx$		sig. → not sig.
Temp.	2.5	2.7	-0.2	$A < B$	.19	2.3	2.6	-0.3	$A < B$	.38	0.1	$\approx$		
Succ.	4.0	3.6	0.4	$A > B$	**.05**	3.8	3.9	-0.1	$A \approx B$	.7	-0.3	Closing	✓	sig. $A > B \rightarrow A \approx B$
Eff.	2.4	3.0	-0.6	$A < B$	**.01**	2.3	2.7	-0.4	$A < B$	.36	-0.2	Closing		sig. → not sig.
Frust.	2.0	2.7	-0.7	$A < B$	**.00**	2.2	2.4	-0.2	$A < B$	.6	-0.5	Closing	✓	sig. → not sig.
2														
Ment.	3.2	3.4	-0.2	$A < B$	.47	3.1	3.5	-0.4	$A < B$	.32	0.2	Widening		
Phys.	2.1	2.1	0.0	$A \approx B$	.69	1.5	2.3	-0.8	$A < B$	**.03**	0.8	Widening	✓	$A \approx B \rightarrow$ sig. $A < B$
Temp.	2.8	2.9	-0.1	$A \approx B$	.30	2.6	3.4	-0.8	$A < B$	.07	0.7	Widening	✓	$A \approx B \rightarrow A < B$
Succ.	3.0	3.2	-0.2	$A < B$	.23	3.2	3.4	-0.2	$A < B$	.38	0.0	$\approx$		
Eff.	3.1	3.2	-0.1	$A \approx B$	.52	2.7	3.4	-0.7	$A < B$	.18	0.6	Widening		$A \approx B \rightarrow A < B$
Frust.	2.6	2.6	0.0	$A \approx B$	.83	2.0	2.7	-0.7	$A < B$	.11	0.7	Widening	✓	$A \approx B \rightarrow A < B$
3														
Ment.	2.6	2.8	-0.2	$A < B$	.68	2.8	3.0	-0.2	$A < B$	.66	0.0	$\approx$		
Phys.	1.9	3.2	-1.3	$A < B$	**.03**	2.0	3.2	-1.2	$A < B$	.13	-0.1	$\approx$		sig. → not sig.
Temp.	2.2	3.1	-0.9	$A < B$	.06	2.4	3.2	-0.8	$A < B$	.27	-0.1	$\approx$		
Succ.	4.3	4.2	0.1	$A \approx B$	1.0	4.4	4.2	0.2	$A > B$	.6	0.1	$\approx$		
Eff.	2.1	3.2	-1.1	$A < B$	.05	1.8	4.0	-2.2	$A < B$	**.01**	1.1	Widening	✓	not sig. $A < B \rightarrow$ sig. $A < B$
Frust.	1.6	2.5	-0.9	$A < B$	.07	1.4	3.6	-2.2	$A < B$	**.02**	1.3	Widening	✓	not sig. $A < B \rightarrow$ sig. $A < B$
4														
Ment.	3.0	2.8	0.2	$A > B$	1.0	3.2	2.4	0.8	$A > B$	.2	0.6	Widening		
Phys.	2.1	2.5	0.4	$A < B$	.19	2.0	1.9	0.1	$A \approx B$	.83	-0.3	Closing		not sig. $A < B \rightarrow A \approx B$
Temp.	2.6	2.6	0.0	$A \approx B$	.84	3.0	2.0	1.0	$A < B$	**.02**	1.0	Widening	✓	$A \approx B \rightarrow$ sig. $A < B$
Succ.	3.7	3.6	0.1	$A \approx B$	.84	3.7	3.7	0.0	$A \approx B$	.77	-0.1	$\approx$		
Eff.	2.9	3.1	-0.2	$A < B$	.38	3.1	2.6	0.5	$A > B$	.28	0.3	Widening	✓	not sig. $A < B \rightarrow$ not sig. $A > B$
Frust.	2.0	1.9	0.1	$A \approx B$	.95	2.4	1.6	0.8	$A > B$	.08	0.7	Widening	✓	$A \approx B \rightarrow$ not sig. $A > B$

In particular, questions remain about how and why these effects occur. In traditional psychology experiments on anchoring effects, participants are given a numerical stimuli (e.g., a high or low value) and then immediately proceed to a subsequent numerical assessment. However, this response immediacy is not typically present when completing NASA-TLX worksheets to assess user interfaces. Instead, participants typically answer a series of questions about a first interface, such as $A_{1-mental}$, $A_{1-physical}$, …, $A_{1-frustration}$, and then proceed to a related series of questions for a second interface. Why, then, should a value for $A_{1-mental}$ serve as an anchor for $B_{2-mental}$ given that a series of values intervened?

We see three candidate explanations, but further work is required to determine their veracity. First, it might be that users simply remember their earlier value, and this memorised value anchors their subsequent one. Second, NASA-TLX values are often recorded on paper worksheets, and these worksheets are sometimes available to participants throughout the experiment. Therefore, when completing an assessment for one category (such as $B_{2-mental}$), the participant might refer back to their earlier value (e.g., $A_{1-mental}$), with this reference refreshing the anchor value. Studies 1, 2 and 3 analysed in this paper all used paper worksheets that permitted referring to earlier responses. Study 4 used computer-based presentation that prohibited back referencing. Finally, it is possible that participants form an overall impression of one interface and that its associated overall

subjective value serves as a form of anchor for their assessment of the subsequent interface (consistent with the halo effects reported by [16]).

Further work to better understand these issues has practical implications for how researchers should gather subjective ratings. For example, if physical worksheets were known to amplify anchoring effects due to eased back-referencing, this would suggest that on-line methods should be preferred, with previous responses suppressed in the data display. However, it seems likely that some memory effects will remain.

Recommendations for Experimenters

In light of our findings, we urge researchers to inspect subjective measures for asymmetric transfer, preferably using the aligned rank transform [40]. When asymmetric transfer is indicated, researchers would be best to revert to between-subjects comparison of data from only the conditions completed first. Furthermore, given the psychological foundations for anticipating asymmetric transfer in subjective measures, between-subjects analysis is probably preferable even when asymmetric transfer is not significantly indicated.

This recommendation to prefer between-subjects treatment has three main problems. First, within-subjects treatment has substantial benefits in increasing test sensitivity, largely because the role of individual differences are mitigated. Adopting between-subjects analysis therefore risks increasing Type II errors. However, given that it is already standard practice to switch to between-subjects treatment when asymmetric transfer is indicated for objective measures, we see little reason for adopting a different practice for subjective measures.

Second, opening the possibility for researchers to analyse within-subjects subjective measures using between-subjects methods may increase opportunities for researchers to 'p-hack' – applying a multitude of tests in the hope that one will show a desirable result, unduly increasing Type I error rate. Preregistration of experimental procedures could mitigate these risks, with researchers stating their analytical methods before conducting the experiment (including their intention to test for asymmetric transfer and revert to between-subjects analyses if indicated) [7].

Third, the recommendation is problematic when experiments involve more than two interface conditions. The analyses reported in this paper only studied experiments in which there were exactly two interfaces. We had datasets from several studies that involved three or more interfaces, but we elected not to analyse them for two reasons. First, anchoring effects might operate in several ways, such as A_1 influencing both B_2 and C_3, or alternatively A_1 might influence B_2, with B_2 then influencing C_3. Second, full counterbalancing across n conditions creates $n!$ different group orders, which will normally cause tiny samples for between-subjects group-based analysis.

Given the frequency of asymmetric transfer in the studies analysed in this paper, and its unknown (but suspected) occurrence in studies with three or more levels, we recommend treating with suspicion any subjective rating data from within-subjects analysis of more than two interfaces.

Other Subjective Measures

Our analysis only examined NASA-TLX responses. The decision to limit the analysis to these measures was intended to minimise extraneous factors and maximise data consistency across studies. However, this creates risks that the findings will not generalise beyond NASA-TLX worksheets. We believe that this is a relatively minor concern because the results conform to a well established and comprehensively evaluated underlying theory of anchoring effects. Regardless, further work validating the occurrence of asymmetric transfer and conformance with anchoring effects is needed for other forms of subjective measure.

Arguably the most promising direction for further study lies in understanding what factors contribute to asymmetric transfer in subjective ratings when the effect of anchoring is minimal or absent. This includes understanding the roles of choice framing, reference points, primacy effects, and other related effects as described in the Background section*.

CONCLUSION

Within-subjects experimental designs are widely used in HCI research, and subjective rating measures are commonly gathered in these experiments. However, within-subjects designs are susceptible to asymmetric transfer effects, and when asymmetric transfer occurs, within-subjects analyses are confounded.

We presented the rationale for suspecting that asymmetric transfer would frequently occur in the subjective data gathered from within-subjects HCI experiments, due to anchoring effects. Yet HCI researchers do not (to our knowledge) inspect subjective data for its troublesome occurrence. Through four analyses of NASA-TLX data from four previously published HCI experiments we confirmed the following: 1. asymmetric transfer was disturbingly common; 2. the data conformed to predicted effects of anchoring; 3. the magnitude of the anchoring effect correlated with the magnitude of the anchoring stimulus; and 4. several of the previously published findings changed when the effects of asymmetric transfer were appropriately addressed through between-subjects treatment. We encourage HCI researchers to be

cautious when inspecting and interpreting subjective measures from within-subjects experiments; asymmetric transfer should be anticipated and tested, with between-subjects treatment preferable when asymmetric transfer is indicated.

REFERENCES

[1] Gediminas Adomavicius, Jesse C. Bockstedt, Shawn P. Curley, and Jingjing Zhang. 2013. Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects. *Information Systems Research* 24, 4 (2013), 956–975. https://doi.org/10.1287/isre.2013.0497 arXiv:https://doi.org/10.1287/isre.2013.0497

[2] Dan Ariely, George Loewenstein, and Drazen Prelec. 2003. "Coherent Arbitrariness": Stable Demand Curves Without Stable Preferences. *The Quarterly Journal of Economics* 118, 1 (2003), 73–106. https://doi.org/10.1162/00335530360535153

[3] Bennet B. Murdock Jr. 1962. The Serial Effect of Free Recall. 64 (11 1962), 482–488.

[4] Roy F. Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. 2001. Bad Is Stronger Than Good. *Review of General Psychology* 5, 4 (2001), 323–370.

[5] Alan Beggs and Kathryn Graddy. 2009. Anchoring Effects: Evidence from Art Auctions. *American Economic Review* 99, 3 (June 2009), 1027–39. https://doi.org/10.1257/aer.99.3.1027

[6] John P. Chin, Virginia A. Diehl, and Kent L. Norman. 1988. Development of an Instrument Measuring User Satisfaction of the Human-computer Interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '88)*. ACM, New York, NY, USA, 213–218. https://doi.org/10.1145/57167.57203

[7] Andy Cockburn, Carl Gutwin, and Alan Dix. 2018. HARK No More: On the Preregistration of CHI Experiments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 141, 12 pages. https://doi.org/10.1145/3173574.3173715

[8] Andy Cockburn, Dion Woolley, Kien Tran Pham Thai, Don Clucas, Simon Hoermann, and Carl Gutwin. 2018. Reducing the Attentional Demands of In-Vehicle Touchscreens with Stencil Overlays. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '18)*. ACM, New York, NY, USA, 33–42. https://doi.org/10.1145/3239060.3239061

[9] J. Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates.

[10] Dan Cosley, Shyong K. Lam, Istvan Albert, Joseph A. Konstan, and John Riedl. 2003. Is Seeing Believing?: How Recommender System Interfaces Affect Users' Opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. ACM, New York, NY, USA, 585–592. https://doi.org/10.1145/642611.642713

[11] Fergus I.M. Craik and Robert S. Lockhart. 1972. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior* 11, 6 (1972), 671 – 684. https://doi.org/10.1016/S0022-5371(72)80001-X

[12] Birte Englich, Thomas Mussweiler, and Fritz Strack. 2006. Playing Dice With Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making. *Personality and Social Psychology Bulletin* 32, 2 (2006), 188–200.

[13] Adrian Furnham and Hua Chu Boo. 2011. A literature review of the anchoring effect. *The Journal of Socio-Economics* 40, 1 (2011), 35 – 42. https://doi.org/10.1016/j.socec.2010.10.008

[14] Carl Gutwin, Andy Cockburn, Joey Scarr, Sylvain Malacria, and Scott C. Olson. 2014. Faster Command Selection on Tablets with FastTap. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 2617–2626. https://doi.org/10.1145/2556288.2557136

[15] SG Hart and LE Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, P Hancock and N Meshkati (Eds.). Elsevier Science, 139–183.

[16] Jan Hartmann, Alistair Sutcliffe, and Antonella De Angeli. 2008. Towards a Theory of User Judgment of Aesthetics and User Interface Quality. *ACM Trans. Comput.-Hum. Interact.* 15, 4, Article 15 (Dec. 2008), 30 pages. https://doi.org/10.1145/1460355.1460357

[17] A. R. Herzog and Jerald G. Bachman. 1981. Effects of Questionnaire Length on Response Quality. *The Public Opinion Quarterly* 45, 4 (1981), 549–559.

[18] James J. Higgins, R. Clifford Blair, and Suleiman Tashtoush. 1990. The Aligned Ranks Transform Procedure. In *Proceedings of the Conference on Applied Statistics in Agriculture*. New Prairie Press, 185–195. https://doi.org/10.4148/2475-7772.1443

[19] Daniel Kahneman and Amos Tversky. 1979. Prospect Theory: An Analysis of Decision under Risk. *Econometrica* 47, 2 (1979), 263–291. http://www.jstor.org/stable/1914185

[20] Gideon B Keren and Jeroen G.W Raaijmakers. 1988. On between-subjects versus within-subjects comparisons in testing utility theory. *Organizational Behavior and Human Decision Processes* 41, 2 (1988), 233 – 247. https://doi.org/10.1016/0749-5978(88)90028-3

[21] Botond Kőszegi and Matthew Rabin. 2006. A Model of Reference-Dependent Preferences. *The Quarterly Journal of Economics* 121, 4 (2006), 1133–1165. https://doi.org/10.1093/qje/121.4.1133

[22] IS Mackenzie. 2013. *Human-Computer Interaction: An Empirical Research Perspective*. Waltham, MA: Morgan Kaufmann.

[23] I. Scott MacKenzie. 2013. *Human-Computer Interaction: An Empirical Research Perspective* (1st ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[24] Sylvain Malacria, Gilles Bailly, Joel Harrison, Andy Cockburn, and Carl Gutwin. 2013. Promoting Hotkey Use Through Rehearsal with ExposeHK. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 573–582. https://doi.org/10.1145/2470654.2470735

[25] Thomas Mussweiler and Fritz Strack. 1999. Hypothesis-Consistent Testing and Semantic Priming in the Anchoring Paradigm: A Selective Accessibility Model. *Journal of Experimental Social Psychology* 35, 2 (1999), 136 – 164. https://doi.org/10.1006/jesp.1998.1364

[26] Galit Nahari and Gershon Ben-Shakhar. 2013. Primacy Effect in Credibility Judgements: The Vulnerability of Verbal Cues to Biased Interpretations: Primacy effect in credibility judgements. *Applied Cognitive Psychology* 27, 2 (2013), 247–255.

[27] J Nielsen and J Levy. 1993. Measuring Usability: Preference versus Performance. *Commun. ACM* 37, 4 (1993), 66–75.

[28] Richard E. Nisbett and Timothy D. Wilson. 1977. The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology* 35, 4 (1977), 250–256.

[29] Richard L. Oliver. 1977. Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation. *Journal of Applied Psychology* 62, 4 (1977), 480–486.

[30] Michael P. A. Page and Dennis Norris. 1998. The Primacy Model: A New Model of Immediate Serial Recall. *Psychological Review* 105, 4 (1998), 761–781.

[31] E. C. Poulton and R. S. Edwards. 1979. Asymmetric transfer in within-subjects experiments on stress interactions. *Ergonomics* 22, 8 (1979), 945–961. https://doi.org/10.1080/00140137908924669 arXiv:https://doi.org/10.1080/00140137908924669 PMID: 527573.

[32] Philip Quinn and Andy Cockburn. 2018. Loss Aversion and Preferences in Interaction. *Human-Computer Interaction* 0, 0 (2018), 1–48. https://doi.org/10.1080/07370024.2018.1433040

arXiv:https://doi.org/10.1080/07370024.2018.1433040

[33] Eeva Raita and Antti Oulasvirta. 2011. Too good to be bad: Favorable product expectations boost subjective usability ratings. *Interacting with Computers* 23, 4 (2011), 363 – 371. https://doi.org/10.1016/j.intcom.2011.04.002 Cognitive Ergonomics for Situated Human-Automation Collaboration.

[34] K. C. Salter and R. F. Fawcett. 1985. A robust and powerful rank test of treatment effects in balanced incomplete block designs. *Communications in Statistics - Simulation and Computation* 14, 4 (1985), 807–828. https://doi.org/10.1080/03610918508812475 arXiv:https://doi.org/10.1080/03610918508812475

[35] Joey Scarr, Andy Cockburn, Carl Gutwin, Andrea Bunt, and Jared E. Cechanowicz. 2014. The Usability of CommandMaps in Realistic Tasks. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 2241–2250. https://doi.org/10.1145/2556288.2556976

[36] Milad Shokouhi, Ryen White, and Emine Yilmaz. 2015. Anchoring and Adjustment in Relevance Estimation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. ACM, New York, NY, USA, 963–966. https://doi.org/10.1145/2766462.2767841

[37] Itamar Simonson and Aimee Drolet. 2004. Anchoring Effects on Consumers' Willingness-to-Pay and Willingness-to-Accept. *Journal of Consumer Research* 31, 3 (2004), 681–690. https://doi.org/10.1086/425103

[38] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131. http://www.jstor.org/stable/1738360

[39] Andre C. Valdez, Martina Ziefle, and Michael Sedlmair. 2018. Priming and Anchoring Effects in Visualization. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 584–594.

[40] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 143–146. https://doi.org/10.1145/1978942.1978963

[41] Youjae Yi. 1990. *A Critical Review of Consumer Satisfaction*. Vol. 4. Chicago, IL: American Marketing Association, 68–123.

[42] Jingjing Zhang. 2011. Anchoring Effects of Recommender Systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)*. ACM, New York, NY, USA, 375–378. https://doi.org/10.1145/2043932.2044010