**QWERTY: We Like Easy Words to Type**
Reviewer: Daniel Casasanto

This paper reports an experiment following up on an idea proposed in Jasmin & Casasanto, 2012 (PBR): Words that are easier to type on the QWERTY keyboard should be, on average, more positive in meaning than words that are harder to type. The authors report that, in a list of 120 words that they constructed, words typed with keystrokes that alternate between hands were significantly more positive than words typed with hand repetitions. However, there was no effect of hand alternation on the list of pseudowords they constructed, and no significant right-side advantage (RSA) in the reported analysis. The authors interpret these data as a challenge to the previously reported QWERTY effect (i.e., RSA predicts valence), and as evidence for a different relationship between QWERTY typing fluency and valence.

I am happy to see the QWERTY effect generating follow-up experiments from other labs, and I want to encourage these authors in any way that I can. Unfortunately, there are serious problems with this paper, of two sorts: Problems with the reported experiment itself, and problems with the authors' critique of the original QWERTY effect. (In some cases, these problems are inseparable.) I'll briefly mention some of the most serious problems, and suggest potential solutions where possible.

### 1. The reported effect, if legitimate, would be complementary to the original QWERTY effect – not a challenge to it.

As the title of this submission suggests, the authors propose that words that can be typed more fluently should be rated as more positive in valence. They suggest that this idea is an alternative to what J&C proposed. But actually, this is the same as what we proposed. This should be abundantly clear from the paper, which contains several statements of this proposal. For example:

"…letters that are easier to type should come to carry more positive associations (and letters that are harder to type more negative associations) and should subtly influence the emotional valence of the words they compose." (J&C, pg. 503).

The present authors' proposal (and the title of their paper) is largely a restatement of J&C's proposal that words that can be typed more fluently should be more positive in valence. There is no disagreement here.

If the effect the authors reported were legitimate (there are several reasons to question whether it is), the Hand Alternation effect would not be a challenge to the QWERTY effect reported by J&C. Rather, it would be a complementary demonstration of the same principle: Words that are easier to type (for various reasons) tend to acquire more positive meanings.

I would be happy to learn that the relationship between hand alternations and word valence that the authors report is significant when properly analyzed, and that it generalizes beyond the small set of words they constructed. I suspect that this is not the case, however, largely because J&C already tested for precisely this relationship – in 5 corpora – and never found it to be significant. (See below.)

**2. Jasmin & Casasanto already conducted analyses of hand alternation and finger repetition.**

The fact that the present authors decided to do an experiment to find out whether the rate of Hand Alternations correlates with word valence suggests they may have overlooked the following paragraphs in J&C's General Discussion:

"[Our] proposal is broadly consistent with previous research showing influences of typing fluency on preference judgments for meaningless letter strings (e.g., Beilock & Holt, 2007; Van den Bergh et al., 1990). However, previous studies have focused on different sources of typing fluency, such as finger repetition. For example, skilled typists prefer pairs of letters typed with different fingers ("f–j") over pairs typed with the same finger during standard touch typing ("f–v"; Beilock & Holt, 2007). **In exploratory analyses, we found no significant relationship between the number of finger repetitions in a word and its valence, nor was there any relationship between valence and the number of hand alternations used when typing a word—for any of the corpora we analyzed.**

**These other sources of typing fluency are orthogonal to the number of right-side and left-side letters in a word, and the effects we report here remain significant when both finger repetition and hand alternation are controlled."** (J&C, pg. 503)

So, J&C already tested for the effect of hand alternation on valence that the authors report here – 5 times – and found that it was not significant in any of the corpora. Furthermore, J&C found that their predicted effect of RSA on valence remained significant when these other sources of fluency were controlled. (Conversely, the present authors never reported any analysis of Hand Alternations controlling for RSA.)

So, why did the present authors (apparently) find an effect of hand alternation on valence in their real-word data set? There are many possible explanations, stemming from an unfortunate number of aspects of the reported experiment that are either highly suspicious or clearly in violation of good practice.

**3. Major problems with the design of the reported experiment and analysis of the data.**

**3a. Authors fail to report that there was no effect of hand alternations (HA) in ANEW.**

There is no effect of the number of hand alternations on word valence in the ANEW corpus. A regression for a relationship between ANEW Word Valence and Number of Hand Alternations in each word yields the following: r-squared=.001; F(1,1032)=0.75, p=.39.

Why didn't the authors run this analysis on ANEW? It would have been easy to see that their hypothesis was not supported.

It appears, however, that they may never have tried an analysis comparable to the ones J&C reported, which includes ALL of the words in each corpus, analyzing their valence as a function of a continuous predictor (i.e., the number of hand alternations in a word can easily be computed and used as a continuous IV). If they had done this analysis, they could have saved themselves a lot of time and effort: HA does not predict Valence in ANEW – or in any of the other corpora J&C analyzed.

**3b. Scant information about how the real word and pseudoword word lists were created.**

Rather than testing their predictor variable in an unbiased manner, in pre-existing corpora (or in nonce words generated according to a clear formula), the authors created their own word lists, saying that ANEW did not have enough words in each of the HA bins they decided to create. (There was no need to create bins: HA can and should be used as a continuous variable.)

So, how did they create their stimuli, such that somehow they found the effect they were looking for in 120 items, when J&C found that it was not present even in corpora with over 2000 items? We can't tell, because we have very little information about the stimuli.

The authors first wrote that they used the ANEW norms "in an effort to replicate J&C 2012", but this is clearly not accurate: ANEW contains 1034 words, each of which has a HA and an RHA that can be easily calculated. If they had wanted to replicate J&C (or provide an equipotent test of a similar fluency-valence hypothesis) they would simply have used all of the ANEW words!

Instead, the created a set of 240 items, writing that they added "76 (31.7%) new words" (pg. 7). This is clearly not accurate. 76 is 31.7% of 240. Yet, half of their 240 stimuli were pseudowords, and there are zero pseudowords in ANEW.

So, that means that by "using ANEW," what the authors really meant was that:

(a) They used ANEW for 37% of their 120 real words.
(b) They used ANEW for 0% of their 120 pseudowords.

Where did the majority of their items come from? How were the mysterious 63% of their real words chosen? How many non-ANEW words were added to each of the cells of their design? How were the 100% of their pseudowords constructed? Were the constructed lists controlled for variables that are known to correlate with valence, like frequency and pronounceability? What precautions (if any) were taken to ensure that words were sampled in an unbiased manner – unbiased, that is, with respect to the predicted effects of both HA and RHA?

Unbiased (and clearly explained) methods for sampling words and constructing pseudowords would be necessary but not sufficient for the reported effects to be interpretable.

**3c. The number of words and range of words chosen are in no way representative.**

Even if the words in the present experiment were selected in some unbiased way, there were 10 words in each cell of the design. Thus, the samples of words cannot possibly be representative of the populations of words with the same HAs in English, to which the authors would hope to generalize these results.

In short, you cannot do corpus linguistics of this sort with a total corpus of 120 words, and cells of 10 words each. Even if effects are statistically significant, they are not generalizable – and indeed, in this study, they did not generalize – not even to the sample of pseudowords with the same HAs.

Furthermore, the authors restricted their sample to words that were 3 or 4 letters long. Why? They wrote that it was because things get "complicated" when you include longer words. Why? Shouldn't the same principles apply? Does "complicated" mean that the authors tried longer words but the effects weren't significant?

A much larger and more varied sample of words would be necessary but not sufficient for the reported effects to be interpretable. (The smallest corpus that J&C analyzed was nearly 10 times larger, and no words were ever excluded from the samples.)

**3d. Raw data never reported, highly suspicious use of exclusions and covariates.**

The authors' hypothesis (though never stated as such) is that those factors that have been shown in previous studies to affect typing fluency should correlate with valence. In their literature review, they correctly identify two factors: Finger repetitions (FR) and hand alternations (i.e., the factors whose non-effects on valence were reported in the passage from J&C in #2, above). Clearly BOTH of these factors should influence valence, according to the authors' hypothesis, and on the basis of the studies they review by Beilock and Holt (2007) and by van de Bergh et al. (1990, etc.) – who reported effects of Finger Repetition on judgments of letter valence: FR should be the main predictor of interest.

It is inexplicable, then, that the authors treat Finger Repetition "as a control variable," and HA as the dependent measure of interest. Why? The clearest prediction on the basis of the pre-QWERTY-effect typing literature is that FR should correlate with word valence. The reader is left wondering whether the authors analyzed the effect of FR, found no correlation, and decided *post hoc* to consider it a "control variable."

Since they set up a clear prediction about the effect of FR, the authors are obligated to report its effects (independently and in combination with other factors). If FR has no effect on word valence – which it did not in J&C's unbiased analyses of 5 much larger corpora – this is a failure to support the authors' hypothesis.

Similarly, among the authors' unfounded attacks on J&C, they say that in our study "typing fluency in terms of both expertise and physical keypress combinations was ignored." (pg. 6) Not true. Actually, J&C discussed both of these factors – see J&C pg. 503.

This oversight is ironic given that the present authors collected typing speed (i.e., expertise) but did not analyze its effects. Their hypothesis clearly suggests that this fluency-related variable should be a predictor of valence (that's their hypothesis, right?) -- but instead, they used it as a "control variable." Why? One wonders whether it's because they tested for effects of individual fluency on word valence and didn't find any.

Overall, the analyses are reported incompletely and opportunistically. Analyses on the raw data are never reported (as they were in J&C); variables that should be predictors are used as covariates (e.g., FR, typing fluency); variables that should be added as covariates or matched across conditions are ignored (e.g., word frequency, letter frequency, pronounceability – factors known to correlate with valence).

Reporting the results of the raw regressions of Valence on FR and of Valence on HA, and then reporting analyses of these regressions with all and only the appropriate control variables, would be necessary but not sufficient to make these data interpretable.

**3e. No itemwise analyses conducted.**

As of 1973, when Herb Clark published his landmark paper on the "items as a fixed effect fallacy," the psycholinguistics community has been aware of a devastating threat to the validity and generalizability of language studies that follows from treating items (e.g., words or pseudowords) as fixed effects. Clark developed a procedure for conducting item-wise analyses (i.e., Combining F1, the subject-wise ANOVA with F2, the itemwise ANOVA, into a valid and generalizable statistic: F-prime (or its approximation, minF-prime). For decades, back when language researchers were still using ANOVAs as the authors do here, top journals like JML required authors to report minF-prime, because subject-wise ANOVAs, alone – which is what the authors report here – are not adequate.

In the past decade (see 2008 JML papers by Harald Baayen, Florian Jaeger), the field has moved away from minF-prime, toward a much more elegant solution to the problem of item-wise generalizability: Mixed-effects linear regression models (as reported in J&C). Without such analyses, it is not possible to generalize the reported results beyond the actual 120 items used!

Appropriate itemwise analyses (i.e., mixed-effect regressions with subjects and items as simultaneous random effects) would be necessary but not sufficient to make these data interpretable.

**3f. It is a serious problem that the pseudoword effect was null.**

The authors find an interaction between Condition (HA bins) and Word Type (word, pseudoword), and therefore analyze the word HA-Valence effects in words and pseudowords separately. They find a main effect of HA for words, but not for pseudowords.

They excuse the null effect in pseudowords, saying that this was because these words had never been typed. Well, the items in J&C's pseudoword corpus had never been typed, either!

J&C showed a highly significant QWERTY effect in 1600 phonotactically legal English pseudowords that had never been typed. Beilock and colleagues and van den Bergh and colleagues found fluency-valence effects in letter bigrams that had probably never been typed in isolation. Therefore, the authors cannot dismiss their null result in pseudowords simply by saying they had never been typed. This null result is a clear failure to support of their hypothesis.

It would seem, on the basis of what's reported, that many of the effects the authors predicted (or should have predicted on the basis of their hypothesis) were null:

Effect of HA on Valence in pseudowords: Not significant
Effect of typing speed on valence in real words: Presumably not significant
Effect of typing speed on valence in pseudowords: Presumably not significant
Effect of FR on valence in real words: Presumably not significant
Effect of FR on valence in pseudowords: Presumably not significant
Effect of HA on Valence in real words, in the raw data, and controlling for all and only standard psycholinguistic variables like frequency and for RHA: Presumably not significant

**4. Major problems with the authors' criticisms of J&C's QWERTY effect.**

**4a. The present study in no way constitutes a replication attempt of the QWERTY effect.**

The authors ran one analysis of RHA predicting valence in their items. It's not clear which items they used (i.e., just the words, or also the pseudowords), but they reported no significant relationship between RHA and valence, and described this as a failure to replicate the QWERTY effect.

The lack of a significant effect of RHA on valence in this analysis is not surprising – chief among the reasons why their analysis was inappropriate is that it is massively underpowered.

Their list of words was 12% of the size of the ANEW corpus (the smallest of the corpora J&C analyzed). Their list of pseudowords was 8% of the size of J&C's pseudoword list.

You can't claim that you fail to replicate an effect if you "attempt" to replicate it with a much, much less powerful design.

**4b. Argument from hearsay.**

The authors' critiques of J&C's QWERTY effect come in two varieties:

(i.) Statements that are simply not true and can be dismissed just by reading the paper (e.g., that we did not consider hand alternations).

(ii.) Hearsay: paraphrasing slanderous things they've read on blogs (slanderous=defamatory and demonstrably false), and citing these blogs as evidence for their own unsupported assertions.

Neither sort of critique makes any positive contribution to scientific discourse. I urge the authors to think carefully about the practice of citing non-peer reviewed blogs as scientific evidence – especially when:

(a) The blogger's claims are contradicted by the peer-reviewed paper they are attacking.

(b) The blogger's claims have been definitively refuted by the authors of the peer-reviewed paper.

(c) The blogger has admitted to making mistakes, and posting incorrect analyses on his blog.

The QWERTY effect makes an excellent test case for the issues of replicability that are so much on psychologists' minds these days. In the case of Bargh's experiments, for example, it is impossible to know why one lab finds an effect and another lab doesn't: There are an infinite number of human factors that could influence the outcome of the experiments. Exact replication is impossible.

By contrast, the QWERTY effect can be replicated exactly – exactly – by anyone with the requisite understanding of inferential statistics.

The authors write that the QWERTY effect reported by J&C was "small," "effectively random" and "can be replicated by a random number generator."

The QWERTY effect is, indeed, small as measured, but:

1. Effect size is orthogonal to the statistical significance and to the theoretical significance of an effect.

2. As we have explained previously, the effect size as reported is for UNAVERAGED data. In the submitted study, as in the great majority of studies in experimental psychology, effect sizes are reported for averaged data. The effect sizes for the SAME DATA SET may differ by orders of magnitude depending on whether the data are averaged (binned) or unaveraged (raw).

3. The authors are confusing effect size with replicability. The QWERTY effect is highly replicable. In the original paper, it was replicated 6 times, in 5 corpora, in 3 languages and novel pseudowords.

> We have replicated it several times subsequently, in other languages and other corpora. (We have also replicated it in single letters – which definitively rules out any concern that the original QWERTY effect was due to hand alternations or finger repetitions.)

Could the QWERTY effect be "replicated by a random number generator"? Absolutely – any statistical effect could be. The question is how often? The answer is:

QWERTY Effect in English ANEW: 4 times out of 100 (p=.04)
QWERTY Effect in Dutch ANEW: 2.5 times out of 100 (p=.025)
QWERTY Effect in Portuguese ANEW: 4 times out of 1,000 (p=.004)
QWERTY Effect in Spanish ANEW: 10 times out of 100 (p=.10)
QWERTY Effect in the combined ANEW Corpora from J&C: 1 time out of 1,000 (p=.001)
QWERTY Effect in AFINN Corpus: 1 time out of 1,000 (p=.001)
QWERTY Effect in J&C's Pseudoword Corpus: 5 times out of 10,000 (p=.0005)
Etcetera, etcetera.

These are the results of correctly conducted permutation analyses, with 10,000 iterations each. Liberman, in the ill-founded critique he posted, conducted a permutation analysis with THREE iterations – completely meaningless. He also sorted the letters incorrectly, reported the wrong statistic (adjusted r2 is not appropriate for a simple regression), and failed to realize that, in fact, his "analysis" supported the QWERTY effect 100%: Because all 3 of his randomizations produced effects that were smaller than the observed QWERTY effect, if we were to interpret his analysis, we would have to conclude that it supports the QWERTY effect unequivocally, with p-value of p=0.000000000000(to infinity). Of course, we would never interpret these blunders as evidence for *or* against the QWERTY effect.

Although blog posts should not be considered part of the scientific record, to clarify their own understanding I would encourage the authors to read J&C's reply to Liberman and Dodds, where we explain why their hastily posted, non-peer-reviewed analyses were incorrect, and how when properly analyzed, the QWERTY effect had been shown to be significant many times, across languages and across corpora.

http://www.casasanto.com/QWERTY.html

If the authors have any questions about these comments, or about how the QWERTY effect can be tested appropriately, I welcome them to contact me.