Behavior Research Methods (2018) 50:2399–2407
https://doi.org/10.3758/s13428-018-1016-9

Reliability of the sliding scale for collecting affective responses to words

C. Imbault[1] · D. Shore[1,2] · V. Kuperman[1,2]

Published online: 25 January 2018

Abstract

Warriner, Shore, Schmidt, Imbault, and Kuperman, Canadian Journal of Experimental Psychology, 71; 71–88 (2017) have recently proposed a slider task in which participants move a manikin on a computer screen toward or further away from a word, and the distance (in pixels) is a measure of the word's valence. Warriner, Shore, Schmidt, Imbault, and Kuperman, Canadian Journal of Experimental Psychology, 71; 71–88 (2017) showed this task to be more valid than the widely used rating task, but they did not examine the reliability of the new methodology. In this study we investigated multiple aspects of this task's reliability. In Experiment 1 (Exps. 1.1–1.6), we showed that the sliding scale has high split-half reliability ($r = .868$ to .931). In Experiment 2, we also showed that the slider task elicits consistent repeated responses both within a single session (Exp. 2: $r = .804$) and across two sessions separated by one week (Exp. 3: $r = .754$). Overall, the slider task, in addition to having high validity, is highly reliable.

Keywords Valence · Arousal · Emotion · Reliability

Words evoke affective responses, which can be indexed through subjective ratings of valence. Norms for these responses are available in English (Bradley & Lang, 1999; Warriner, Kuperman, & Brysbaert, 2013), French (Bonin et al., 2003; Monnier & Syssau, 2014), Spanish (Hinojosa et al., 2016; Redondo, Fraga, Padrón, & Comesaña, 2007; Stadthagen-Gonzalez, Imbault, Pérez Sánchez, & Brysbaert, 2017), Dutch (Moors et al., 2013), and other languages. All of these studies, dating back to 1999 (Bradley & Lang, 1999), have one method in common: They present words in isolation and instruct participants to evaluate the words' valence (from *negative* to *positive*) and arousal (from *calm* to *excited*) on a rating scale from 1 to 9 (or, in some cases, 1 to 7). A number of criticisms of this methodology led us to develop a new methodology (Warriner, Shore, Schmidt, Imbault, & Kuperman, 2017), summarized below. In the present study, we evaluated the reliability of this sliding-scale methodology.

Criticisms of the popular rating scale fall into two basic categories (Warriner et al., 2017; Westbury, Keith, Briesemeister, Hofmann, & Jacobs, 2015). The first class of criticism concerns the nature of the data collected. A rating scale with 9 (or 7) individual points produces an ordinal measure, whereas an interval measure is preferable for most statistical tests. In fact, most articles have treated this ordinal scale as an interval scale for statistical analyses, which can lead to violations of assumptions. Related to this general criticism, the typical rating scale does not allow for fine-grained output—observers can only provide integer responses, which can fail to capture the subtle effects of emotion evoked by a word or individual variability in affective behavior.

The second class of criticism concerns the necessity, within a typical rating study, to anchor the affective evaluation. Depending on the words chosen for the anchors, observers can be biased to overemphasize some words and diminish the impact of others (Westbury et al., 2015). Typically the words chosen (e.g., *pleasant* and *unpleasant*) are too mild for words that raters encounter in these studies, such as *rapist*. Inappropriate anchoring can lead to drift across the study once extreme words are encountered. Ideally, the scale should allow observers to maximize the range of values used without relying on the specific words presented.

To counter these concerns, we developed the *slider* task (Warriner et al., 2017). A humanoid manikin is placed in the center of a vertical line, with a word at the top or

✉ C. Imbault
imbaulcl@mcmaster.ca

[1] Department of Linguistics & Languages, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada

[2] Department of Psychology, Neuroscience & Behaviour, McMaster University, Hamilton, Ontario, Canada

bottom of the line. Participants, who are instructed that the manikin represents themselves, can slide the manikin as close to or far away from the word as they prefer. The distance from the word represents the participants' affective response to the word; a greater distance indicates lower valence (more negative), and a smaller distance indicates higher valence (more positive). Distance is a continuous interval-scale variable, which is in practice discretized into a number of pixels[1] on a computer screen, just as the continuous variable of time is discretized into fractions of seconds by chronometric instruments. Warriner et al. (2017) found that distance was negatively correlated with word valence in a variety of populations (undergraduates in lab, $r = -.62$; adults online, $r = -.58$). Additionally, Warriner et al. (2017) found that the slider task is sensitive to individual differences; for instance, those who are shyer tend to position the manikin farther from all stimuli than do those who are less shy (by 56 pixels, or 10% of the scale). As well, the slider is sensitive to gender differences: That is, females position the manikin closer to words that are rated as being more positive by female than by male raters, and the same is true for males. The range of 600 pixels that we utilize in the slider task also allows researchers to capture subtle individual differences that are lost in a smaller, 9-point ordinal scale (e.g., the subtle tendency of relatively sociable individuals to keep a shorter distance to all words; Warriner et al., 2017). These findings are in line with an earlier proposal to move from discrete rating scales to a continuous slider by Albaum, Best, and Hawkins (1981). They argued that a discrete rating scale and a continuous scale produce similar aggregate data, but that a continuous scale allows for greater discrimination at the individual level.

Finally, this task has no anchors and removes the mention of valence from the task instructions. By tapping into implicit approach–avoidance tendencies, the task avoids explicit linkages to artificial valence terminology (e.g., *pleasant* and *unpleasant*). The lack of traditional semantic labels alluding to valence may alter the psychological construct measured by the slider task, causing participants to tap into approach–avoidance behavior instead of producing affective responses (see the General Discussion below, as well as the Future Directions section in Warriner et al., 2017). An argument can also be made that the lack of anchors may cause participants to be confused and not to perform the task as intended. Although possible, it is unlikely that participants are confused or would engage in a different behavioral pattern. Warriner et al. (2017) administered the slider task with and without anchors and found that there was no difference in the functional relationship between valence and distance. In sum, the slider task provides a new method of collecting affective ratings that is more valid than past methods. The utility of a similar affective slider method for measuring the valence and arousal of pictures taken from the International Affective Picture System (IAPS) has been demonstrated by Betella and Verschure (2016).

Given the validity of the task (Warriner et al., 2017), we expect many researchers and clinicians to be interested in the assessments of emotion that come from this task; however, before we can evaluate individual and group differences in emotional responses, we must ask about the method's reliability. The present article evaluates the reliability of the slider task through several converging methods: In Experiment 1 we utilized previously collected data and applies a split-half analysis (cf. MacLeod et al., 2010); in Experiment 2 we collected new data in a repeated measures design within a single data collection session; and Experiment 3 was based on a standard test–retest design from two sessions separated by a week. In all cases, the measure of interest was the slope of the best-fitting regression line between distance from the word (cf. Warriner et al., 2017) as the dependent measure, and the normed rated valence of the word (cf. Warriner et al., 2013) as the independent variable. To be specific, we assessed to what extent the slope of this line from one sample predicted the slope from a second sample within the same individual. In the split-half analysis, the two samples came from the same session (Exp. 1); in the repeated measures design, observers rated the same words in two separate blocks of trials in one session (Exp. 2); and in the test–retest design, the same words were rated on two separate occasions separated by one week (Exp. 3). We performed an additional analysis in Experiments 2 and 3 to remove the effect of valence and instead measure the reliability of distance. We assessed to what extent the distance from the word in one sample predicted the distance from the word in a second sample within the same individual.

Experiment 1: Split-half analysis

The split-half analysis utilized previously collected data (Warriner et al., 2017) to estimate the reliability of the slider task. We combined the standard split-half analysis with a resampling technique in order to enhance the robustness of our estimate and provide confidence intervals on our reliability estimates (cf. MacLeod et al., 2010). Each experiment presented the same 250 words drawn from separate quintiles of valence; these were 50 words

[1] We thank our reviewers for bringing double-density displays to our attention. In future studies, we will refer to "steps" as our dependent variable of interest, rather than "pixels," which may be an inaccurate measurement on double-density displays.

rated for each of five different valence subranges. To determine split-half reliability, we randomly selected two samples of 20 words from each quintile without replacement, creating two groups of 100 words. For each group and each participant, we estimated a regression slope for distance as predicted by valence; two regression slopes were thus generated for each participant. Reliability was estimated by examining the correlation of these two slopes across participants. This sampling was conducted 10,000 times for each of the five experiments, four of which are reported in Warriner et al. (2017); the fifth data set (Experiment 1.3) came from the first session of the present article's Experiment 3 (these analyses are labeled here Exps. 1.1–1.5). The data from the current article's Experiment 2 were not included in the split-half analysis because Experiment 2 only involved half of the 250 words used in Experiments 1.1–1.5. These five experiments varied in their experimental settings, task instructions, and participant demographics. In Experiment 1.6, we analyzed the combined data from all five experiments, weighting each experiment by the number of participants.

Method

Participants and procedure All experiments used a similar set of instructions and similar numbers of participants (see Table 1 for specifics from each experiment; see Warriner et al., 2017, for the detailed methods). Each participant was seated in front of a computer monitor with a screen resolution of 1,024 × 768 placed approximately 57 cm away (Exps. 1.4 and 1.5 used an online version of the task—the specifics of the monitor and computer thus varied across participants). Following a central fixation point, a humanoid manikin was presented at the center of the computer monitor along a

Table 1 A summary of the participants, procedure, and critical results (correlations between valence and distance, as well as split-half reliability) of Experiments 1.1–1.5

	Participants	Procedure	Split Half Results	Spearman Brown Correction
1.1	43 participants (35 female), ranging from 17 to 25 years of age (*M* = 19.07, *SD* = 1.98)	Participants at McMaster University took part in the study for partial course credit. No more than ten participants at a time completed the study in a computer lab on campus. The data were taken from Warriner et al.'s (2017) Experiment 1.	ρ = .873, 95% CI [.800, .926]	ρ = .932, 95% CI [.889, .962]
1.2	30 participants (all female), ranging = in age from 18 to 21 (*M* = 19, *SD* = 1.02)	Participants at McMaster University took part in the study for partial course credit. No more than ten participants at a time completed the study in a computer lab on campus. Prior to the start of the experiment, participants were given four personality questionnaires to fill out. The data were taken from Warriner et al.'s (2017) Experiment 2.	ρ = .868, 95% CI [.786, .929]	ρ = .929, 95% CI [.880, .963]
1.3	36 participants (33 female) ranging in age from 18 to 27 (*M* = 18.69, *SD* = 1.59)	Participants at McMaster University took part in the study for partial course credit. No more than ten participants at a time completed the study in a computer lab on campus. The data were taken from the first session of the present Experiment 3.	ρ = .883, 95% CI [.812, .934]	ρ = .938, 95% CI [.896, .966]
1.4	32 participants (13 female), ranging in age from 19 to 55 (*M* = 34.06, *SD* = 8.68)	An ad was posted on Amazon Mechanical Turk for participants to complete the study from their home for monetary compensation. All participants were based in the USA. These data were taken from Warriner et al.'s (2017) Experiment 3.	ρ = .931, 95% CI [.886, .963]	ρ = .964, 95% CI [.939, .981]
1.5	36 participants (13 female) ranging in age from 21 to 60 (*M* = 34.42, *SD* = 8.95)	An ad was posted on Amazon Mechanical Turk for participants to complete the study from their home for monetary compensation. All participants were based in the USA. Anchors were removed from the instructions given to participants. The data were taken from Warriner et al.'s (2017) Experiment 4.	ρ = .898, 95% CI [.831, .945]	ρ = .946, 95% CI [.908, .972]
1.6	177 participants (124 female), ranging in age from 17 to 60 (*M* = 24.81, *SD* = 9.31)	The data from Experiments 1.1 through 1.5 were compiled to create one data set.	ρ = .891, 95% CI [.859, .917]	ρ = .942, 95% CI [.924, .957]

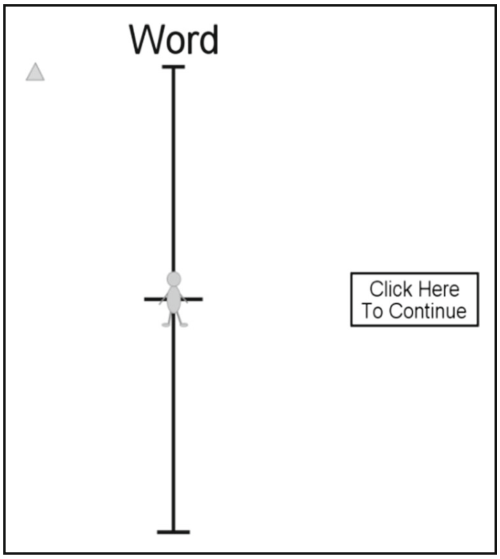

Fig. 1 The slider scale: a humanoid manikin and word for each trial

line with a single word at the top or bottom of the screen (see Fig. 1). Participants moved the manikin up or down to position it at their preferred location as close to, or as far away from, the word as they wished. Experiments 1.1, 1.2, 1.3, and 1.4 all had the same instructions, which read:

> [On a] screen, you will see a word at the top of the screen with a vertical line below it. There will be a person in the centre of that line. The person represents you. Your job is to assess how close you would like to be to the word and communicate that by clicking a point on the line to position the person (you). For example, if the word was DISASTER, you'd probably want to be far away and would click somewhere on the line far away from the word. But if the word was TRIUMPH you might want to be close and would place the manikin somewhere on the line really close to the word.

Experiment 1.5 had a slightly modified set of instructions, in which such anchor words as *disaster* or *triumph* were not mentioned. Those instructions read:

> On each of the following screens, you will first see a plus sign in the centre. That's to center the mouse for the next screen. Click on the plus and you will see a word either at the top or the bottom with a vertical line below or above it. There will be a person in the centre of that line. The person represents you. You can move "yourself" closer to or further away from the word. Position yourself where you prefer to be.

Stimuli The word set for each experiment was the same. It consisted of 250 monosyllabic words chosen from a set of 13,763 words that had previously been rated for valence and arousal (Warriner et al., 2013). The words were divided into 25 bins (5 quintiles of valence × 5 quintiles of arousal), with ten words drawn randomly from each bin. Thus, there were 50 words at each of five levels of valence, which varied in their arousal levels. This ensured that valence and arousal were not correlated ($p = -.019$), and thus in what follows we only measured any effects of valence and not of arousal. The mean word length was 4.4 characters, and the mean natural-log SUBTLEX frequency was 6.3 (Brysbaert & New, 2009). We used natural-log frequency because frequency is represented exponentially, so the natural log makes this distribution closer to normal and easier to interpret.

Results and discussion

The primary dependent variable for all experiments was the distance, measured in pixels, between the position of the manikin and the presented word when the participant pressed the Submit button [range = 1 (*closest to the word*)–600 (*farthest from the word*)]. The participants were able to move the manikin as many times as they wanted before clicking Submit—only the final location of the manikin was used in our data analysis. The independent variable of interest for all experiments presented here was the word's valence rating (Warriner et al., 2013), which varied between 1 (*very unhappy*) and 9 (*very happy*); each word was rated by at least 20 raters.

To assess the split-half reliability of each experiment, we separated the 250 words in the stimulus list into five bins of valence, broken down by quantiles. We randomly selected 20 words (from the approximately 50 words), without replacement, from each bin twice for each participant, to construct two separate groups of 100 words (20 words × 5 bins). In each group, we computed the slope (i.e., the beta coefficient) of the respective regression line between valence and distance. This resulted in two slope coefficients for each participant. We then found the correlation for these two slope estimates across all participants. We repeated this process 10,000 times, which resulted in 10,000 correlations. The mean of these 10,000 correlations provided our estimate of the split-half reliability; the distribution of correlations allowed us to estimate confidence intervals. We used R version 3.01 (R Development Core Team, 2013) to perform our statistical analyses in this and subsequent experiments.

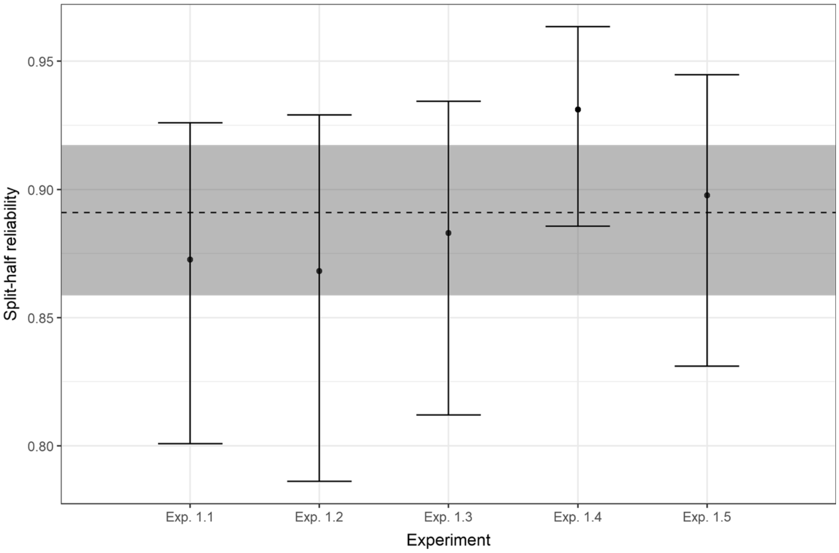

Fig. 2 Estimated split-half reliabilities of Experiments 1.1–1.5, with 95% confidence intervals, mapped on the *y*-axis. The gray shaded area depicts the 95% confidence interval of the weighted mean of the reliability estimates

The split-half reliability for each of Experiments 1.1–1.5 ranged from .868 to .931 (see Table 1 and Fig. 2), which can be considered a very high level of reliability. Experiment 1.6 combined the data from all five experiments and weighted each experiment by the number of participants. This analysis produced a reliability of .891 [95% CI = .859–.917]. Thus, this new method of collecting affective ratings of words provides a reliable measure of valence. This was found across multiple experiments with different participant pools, recruitment methods, and experimental settings.

Experiment 2: Within-session repeated measures reliability

Although the split-half analysis provides a good estimate of reliability, there are some challenges in the present case. First, different words were used across the two samples for each participant in each simulated set. Second, the data for each participant were taken from the same experimental session, and the participant was thus in the same mental state—if we want to conclude that this task provides a stable measure of valence, we would need to test participants multiple times. Specifically, we would need to assess the ability of the task to elicit similar responses to multiple presentations of the same stimuli. The present experiment used a test–retest paradigm, in which the same words were presented twice within a 1-h experimental session. Both blocks, which were completed sequentially without any substantial break, contained identical stimuli presented in pseudorandom order. To test the reliability of distance as a metric of valence, we evaluated the correlation between two slopes (one from each block) of the best-fitting regression line, where valence was used to predict distance from the word (Analysis 1). Unlike Experiments 1.1–1.5, we used all words in the two blocks to compute the regression slope estimates. To test the reliability of distance to a word as a behavioral outcome (regardless of valence), we evaluated the correlation between responses to any given word produced by each participant in Block 1 versus Block 2 (Analysis 2).

Method

Participants Sixty-four undergraduate students at McMaster University in Hamilton, Ontario, participated in this experiment for partial course credit. The data from 12 participants were removed (seven did not make a response on more than 25% of the trials [i.e., on 38 trials], and an additional five were not native speakers of English). The remaining 52 participants (43 women, nine men; 48 right-handed) ranged in age from 17 to 22 (M = 18.75, SD = 1.17).

Affective stimuli We randomly selected half of the words (125 words) from Warriner et al. (2017) Experiment 1: For the word selection criteria, see Experiment 1. All participants in this experiment saw the same 125 words.

Procedure Participants were tested in groups of ten or fewer in a computer lab. Each participant was seated in front of a monitor with a screen resolution of 1,024 × 768. After completing a set of demographic questions (including age, sex, handedness, and education), participants were instructed to complete the slider task. The task began with a fixation cross centered on the screen; each trial started by clicking on the cross. The fixation cross was replaced with a humanoid manikin centered on a vertical line in the center of the computer screen; the word for that trial was randomly presented at either the top or bottom of the vertical line, and participants were instructed to slide the manikin (or to click a destination on the line) however close to or far away from the word as they wanted the manikin to be. The instructions were the same as in the present Experiments 1.1–1.4. After the participants had moved the manikin to its final position, they clicked the Submit button, located to the right of the slider (see Fig. 1). The experiment was programmed using the Experiment Builder software (SR Research, Kanata, ON, Canada).

Participants completed five practice trials and then were asked if they had any questions before the experiment proceeded. Each participant saw each of the 125 words twice, once in the first half of the experiment (Block 1) and once in the second half (Block 2). The word order was randomized separately in the two blocks of the experiment. The participants were not made aware of the separate blocks and were not told the test–retest purpose of the study. Because of a programming error, some participants saw an additional five words for a third time at the end of the experiment. These additional responses were removed from the data analysis.

Results and discussion

Analysis 1 The variables used in this analysis were the same as those used in Experiment 1, with the manikin's distance from the word as the dependent variable and word valence as the independent variable.

To compare individual performance between experimental blocks, we first estimated the effect of valence on the distance of the manikin from the word for each participant and each half of the experiment. Specifically, for each participant we fitted two ordinary linear regression models (one for the 125 words in each half) estimating the effect of valence on the manikin's distance from the word. We operationalized the effect as the slope (the beta coefficient) of the regression line. Since the words presented in both blocks were identical (though ordered differently), reliable performance would elicit similar slopes in an individual throughout the experiment. This step generated two sets of 52 slope coefficients (corresponding to the 52 participants and the two blocks). We then calculated the correlation between the two sets of regression coefficients as a measure of the test–retest reliability within one experimental session.

The distance of the manikin from the word was negatively correlated with the valence of the word in both blocks (first block: Pearson's r = – .572, df = 6,498, 95% CI [– .587, – .555], $p < .001$; second block: Pearson's r = – .570, df = 6,498, 95% CI [– .586, – .553], $p < .001$) as well as in the whole experiment (Pearson's r = – .570, df = 12,998, 95% CI [– .582, – .559], $p < .001$). As we had observed in our previous work (Exp. 1; Warriner et al., 2017), participants moved the manikin closer to positive words and farther away from negative words.

The correlation between the two sets of slope coefficients was high (r = .804, df = 50, 95% CI [.680, .883], $p < .001$). This demonstrates that valence affected participants' performance in the slider task similarly in the first and second halves of the experiment and also elicited similar responses to the same stimuli presented a second time. The slider task has a high test–retest reliability within one experimental session.

Analysis 2 The variables used in this analysis were the manikin's distance from the word as the dependent variable and block number as the independent variable.

We compared the distance from the word within participants across the two blocks. Specifically, for each word we generated two lists of distance values for Block 1 and Block 2, respectively: For instance, the distance to the word *chore* produced by Participant 1 in Block 1 would make the first element in List 1, and distance to the same word produced by the same participant in Block 2 would make the first element in List 2. Thus, each list contained 52 elements (corresponding to 52 participants), repeated for each word. We correlated the two lists for each word, with the correlation coefficient as a measure of how reliably that word was rated within participants across blocks. The second step was to evaluate the central tendency and dispersion of correlation strength across words. The mean correlation, or the reliability of the distance measurement, within a single session was moderately high (r_{mean} = .603, SD = .155, CI [.297, .867]). This demonstrates that the participants responded reliably, with similar distances in the first and second blocks within a single session.

Experiment 3: Across-session test–retest reliability

In Experiment 2, we showed that participants' performance was replicable within one test–retest session. Another aspect of reliability is whether the task elicits similar responses to the same stimuli over time. We presented the same slider task to participants twice, in experimental sessions separated by a week. The task was expected to produce similar results over the two sessions.

Method

Participants Forty-five undergraduate students at McMaster University participated in this study for partial course credit. None of the participants had participated in similar studies prior to completing this study. Forty-two of the participants attended the second, retest part of the experiment. The remaining three participants were excluded from the data analysis. Three additional participants were removed from analysis because they did not respond on more than 25% of the trials (i.e., on at least 62 trials). An additional three participants were removed from the analysis due to not being native speakers of English. The remaining 36 native English speakers (33 female, three male) ranged in age from 18 to 27 years (M = 18.69, SD = 1.58).

Affective stimuli The stimuli used in this experiment were the same list of 250 words from Warriner et al. (2017), described in Experiment 1.

Procedure The procedure for this experiment was nearly identical to that of Experiment 2. Participants signed up to take part in a two-part study. The second session took place a week after the first. The instructions given to participants were the exact same as those given in Experiment 2. Participants were first shown five practice words, and then were presented with 250 experimental trials. The 250 words were presented in one block and were randomly ordered. The participants were given the same set of instructions and the same words the second time they came in for the study, though the order of words in each session was randomized differently. Participants were not told that the words presented would be the same as in the first session, and they were not informed of the purpose of the study. Because of a programming error, some participants saw some words for a third or fourth time at the end of the experiment. These additional responses were removed from the data analysis.

Results and discussion

Analysis 1 The variables used in this analysis were the same as those used in Experiment 1 and Analysis 1 of Experiment 2, with the manikin's distance from the word as the dependent variable and word valence as the independent variable.

To measure the reliability of individual performance in the slider task over time, we first estimated the effect of valence on the distance of the manikin from the word for each participant and each half of the experiment, using ordinary linear regression models (see Exp. 2). This step generated two sets of 36 slope coefficients (corresponding to 36 participants and the two blocks). We then calculated the correlation between the two sets of regression coefficients as a measure of the test–retest reliability within the two experimental sessions separated by one week.

The distance of the manikin was negatively correlated with the valence of the word for both sessions (first session: Pearson's r = – .559, df = 9,178, 95% CI [– .573, – .544], p < .001; second session: Pearson's r = – .532, df = 9,178, 95% CI [– .546, – .517], p < .001) and for the two sessions together (Pearson's r = – .545, df = 18,358, 95% CI [– .555, – .535], p < .001). Participants moved the manikin closer to positive words and farther away from negative words.

We calculated the effect of valence on distance for each participant and each experimental session. The correlation between the two sets of slope coefficients was strong and significant (Pearson's r = .754, df = 34, 95% CI [.566, .868], p < .001). This demonstrates that the valence of the same set of stimuli affected participants' performance in the slider task similarly in the first and second experimental sessions. The slider task thus has high test–retest reliability over time.

Analysis 2 The variables and statistical analysis were identical to those in Analysis 2 of Experiment 2; the manikin's distance from the word was the dependent variable, and block number was the independent variable. The only exception was that the difference between Block 1 and Block 2 in Experiment 2 are now represented as Session 1 and Session 2.

The mean correlation, or the reliability of the distance measurements, across the two experimental sessions was moderate (r_{mean} = .436, SD = .201, CI [.047, .805]). This demonstrates that the participants responded reliably, with similar distances in the first and second experimental sessions.

The reliability in Experiment 3 was, unsurprisingly, lower than that in Experiment 2 (r = .754 vs. .804, respectively). This was due to the difference in the length of time between the retest sessions; Experiment 2 had no time between sessions, whereas Experiment 3 had a week between the sessions. The same decrease was found in the reliability of the distance measurement, in which a separation by a week (r_{mean} = .436) caused more variability in responses than was found in the same-day session (r_{mean} = .603).

General discussion

The slider task provides a reliable, valid, and easily reproducible way to capture human affective judgments of words presented in isolation. We hypothesized that this conclusion could be extended to include other types of stimuli, such as pictures, sounds, or linguistic phrases. As we argued in Warriner et al. (2017), the slider task addresses numerous criticisms relating to the validity of rating-scale affective judgments. For one thing, the slider task removes any emotional anchors from the instructions that may potentially affect participants' ratings (see Westbury et al., 2015). Warriner et al.

(2017) demonstrated that the functional relationship between valence and distance was not affected by the presence or absence of anchors. The present study has additionally demonstrated that there is no difference in the reliability of the task (see Exps. 1.1–1.4 with anchors, and Exp. 1.5 without anchors). Critically, the slider task also allows for a more fine-grained measure of affective judgments and for robust detection of subtle individual differences in response patterns on an interval scale. A question that was left unanswered in Warriner et al. (2017) was whether the slider task and its measurement of individual differences is reliable.

This article confirms that the slider task performed well on many psychometric reliability measures. In Experiment 1, we demonstrated that the slider task has a high split-half reliability across multiple studies. Participants responded to a word's valence with a high degree of similarity across randomly divided halves of experiments. In Experiment 2, we further confirmed that the slider task is reliable across an hour-long session in which a set of words were presented twice. Participants responded similarly to a word's valence in both the first and second halves of the session. In Experiment 3, we showed that the slider task is reliable over time. Participants were affected by valence similarly within two sessions that presented the same words separated by an interval of a week. We also observed in Analyses 2 of Experiments 2 and 3 that the reliability of producing a distance value is consistently lower than the reliability of that distance value as an index of valence. In other words, over two sessions participants were less consistent in how they approached or withdrew from a given word, but in both sessions those responses were driven by the word's valence to a greater degree and with greater consistency. This suggests that the absolute values of the distance to words varying in valence are less reliable (and theoretically less informative) than the values of distance relative to the word's valence.

We believe that these reliability results, along with the findings from Warriner et al. (2017), show that a sliding scale is a useful tool for collecting affective responses to words, in that it is both valid and reliable. The slider scale shows added utility over a typical discrete rating scale. Both scales produce similar aggregate data, but the continuous slider scale allows for more detailed responses that are crucial when studying individual differences. As well, unlike a rating scale, a sliding scale does not rely on anchor terms that are related to valence, a problem that was discussed in Westbury, Keith, Briesemeister, Hofmann, and Jacobs (2015). Additionally, affective ratings are often used in analyses as interval data, even though they collect ordinal data. The data obtained from the sliding scale are interval by their nature.

One possible limitation of the slider task is discussed at length in Warriner et al. (2017): The task may tap valence, or a related but different psychological construct such as approach–avoidance behavior, or both. Valence is known to be strongly linked to the motivational systems that drive appetitive and aversive responses, to such a degree that disentangling these constructs presents a considerable difficulty (see, among others, Carver & White, 1994; Lang, 1995). The present method does not enable us to partial out independent contributions of different motivational systems to individual affective responses. Although the slider scale is demonstrably a reliable tool for quantifying valence, the question of its link to the approach–avoidance behavior will need to be studied further.

Moreover, it is important to note that we are looking at only one of the two dimensions of affect (Russell, 1980)—that is, at valence but not arousal. Valence is the measurement of negativity to positivity, whereas arousal is the measurement of calmness to excitedness (Russell, 1980). The wording of our sliding scale only taps into measuring valence, whereas other slider scales are able to capture arousal as well (Betella & Verschure, 2016).

In sum, multiple methods confirm that the slider task is reliable within participants. Additionally, our split-half analysis of a combined data set (see Exp. 1.6) showed the task's reliability between participants. We conclude that the task has the high degree of reliability that is essential not only for the overall utility of the task in measuring affective responses, but also for the ability to detect subtle patterns of individual variability.

Author note This work was supported by a Patterson-Wilson Ontario Graduate Scholarship to C.I. V.K.'s contribution was supported in part by the Natural Sciences and Engineering Research Council or Canada (NSERC), Discovery Grant 402395-2012; Grant NIH R01 HD 073288 (PI Julie A. Van Dyke); SSHRC Partnership Training Grant 895-2016-1008 (PI Gary Libben); an Early Research Award from the Ontario Ministry of Research and Innovation; and a Canada Research Chair (Tier 2) award. D.I.S. was also supported by an NSERC Discovery Grant.

References

Albaum, G., Best, R., & Hawkins, D. E. L. (1981). Continuous vs. discrete semantic differential rating scales. *Psychological Reports*, *49*, 83–86.

Betella, A., & Verschure, P. F. M. J. (2016). The affective slider: A digital self-assessment scale for the measurement of human emotions. *PLoS ONE*, *11*, e148037:1–11. https://doi.org/10.1371/journal.pone.0148037

Bonin, P., Méot, A., Aubert, L., Malardier, N., Niedenthal, P., & Capelle-Toczek, M.-C. (2003). Normes de concrétude, de valeur d'imagerie, de fréquence subjective et de valence émotionnelle pour 866 mots [Concreteness, imagery value, subjective frequency, and emotional valence norms for 866 words]. *L'Année Psychologique*, *104*, 655–694. https://doi.org/10.3406/psy.2003.29658

Bradley, M. M., & Lang, P. J. (1999). *Affective Norms for English Words (ANEW): Stimuli, instruction manual and affective ratings* (Technical Report No. C-1). Gainesville, FL: University of Florida, NIMH Center for Research in Psychophysiology.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of present word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–990. https://doi.org/10.3758/BRM.41.4.977

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation and affective responses to impending reward and punishment: The BIS/BAS Scales. *Journal of Personality and Social Psychology*, *67*, 319–333. https://doi.org/10.1037/0022-3514.67.2.319

Hinojosa, J. A., Martínez-García, N., Villalba-García, C., Fernández-Folgueiras, U., Sánchez-Carmona, A., Pozo, M. A., & Montoro, P. R. (2016). Affective norms of 875 Spanish words for five discrete emotional categories and two emotional dimensions. *Behavior Research Methods*, *48*, 272–284. https://doi.org/10.3758/s13428-015-0572-5

Lang, P. J. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist*, *50*, 372–385. https://doi.org/10.1037/0003-066X.50.5.372

MacLeod, J. W., Lawrence, M. A., McConnell, M. M., Eskes, G. A., Klein, R. M., & Shore, D. I. (2010). Appraising the ANT: Psychometric and theoretical considerations of the Attention Network Test. *Neuropsychology*, *24*, 637–651. https://doi.org/10.1037/a0019803

Monnier, C., & Syssau, A. (2014). Affective Norms for French Words (FAN). *Behavior Research Methods*, *46*, 1128–1137. https://doi.org/10.3758/s13428-013-0431-1

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., van Schie, K., Van Harmelen, A.-L., … Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods*, *45*, 169–177. https://doi.org/10.3758/s13428-012-0243-8

R Development Core Team. (2013). R: A language and environment for statistical computing (Version 3.01). Vienna: R Foundation for Statistical Computing. Retrieved from www.R-project.org

Redondo, J., Fraga, I., Padrón, I., & Comesaña, M. (2007). The Spanish adaptation of ANEW (Affective Norms for English Words). *Behavior Research Methods*, *39*, 600–605. https://doi.org/10.3758/BF03193031

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*, 1161–1178. https://doi.org/10.1037/h0077714

Stadthagen-Gonzalez, H., Imbault, C., Pérez Sánchez, M. A., & Brysbaert, M. (2017). Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods*, *49*, 111–123. https://doi.org/10.3758/s13428-015-0700-2

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*, 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

Warriner, A. B., Shore, D. I., Schmidt, L. A., Imbault, C. L., & Kuperman, V. (2017). Sliding into happiness: A new tool for measuring affective responses to words. *Canadian Journal of Experimental Psychology*, *71*, 71–88. https://doi.org/10.1037/cep0000112

Westbury, C., Keith, J., Briesemeister, B. B., Hofmann, M. J., & Jacobs, A. M. (2015). Avoid violence, rioting, and outrage; approach celebration, delight, and strength: Using large text corpora to compute valence, arousal, and the basic emotions. *Quarterly Journal of Experimental Psychology*, *68*, 1599–1622. https://doi.org/10.1080/17470218.2014.970204