i-PERCEPTION

a Pion publication

*i-Perception* (2013) volume 4, pages 365–379

dx.doi.org/10.1068/i0586 ISSN 2041-6695 perceptionweb.com/i-perception

# Fast lemons and sour boulders: Testing crossmodal correspondences using an internet-based testing methodology

Andy T. Woods
Xperiment, Lausanne, Switzerland; e-mail: Andy.Woods@xperiment.mobi

Charles Spence
Crossmodal Research Laboratory, Department of Experimental Psychology, Oxford University, UK;
e-mail: Charles.Spence@psy.ox.ac.uk

Natalie Butcher
Faculty of Health and Life Sciences, York St John University, UK; e-mail: N.Butcher@yorksj.ac.uk

Ophelia Deroy
Centre for the Study of the Senses, School of Advanced Study, University of London, London, UK;
e-mail: ophelia.deroy@sas.ac.uk

Received 24 January 2013, in revised form 8 July 2013; published online 29 July 2013.

**Abstract.** According to a popular family of hypotheses, crossmodal matches between distinct features hold because they correspond to the same polarity on several conceptual dimensions (such as active–passive, good–bad, etc.) that can be identified using the semantic differential technique. The main problem here resides in turning this hypothesis into testable empirical predictions. In the present study, we outline a series of plausible consequences of the hypothesis and test a variety of well-established and previously untested crossmodal correspondences by means of a novel internet-based testing methodology. The results highlight that the semantic hypothesis cannot easily explain differences in the prevalence of crossmodal associations built on the same semantic pattern (fast lemons, slow prunes, sour boulders, heavy red); furthermore, the semantic hypothesis only minimally predicts what happens when the semantic dimensions and polarities that are supposed to drive such crossmodal associations are made more salient (e.g., by adding emotional cues that ought to make the good/bad dimension more salient); finally, the semantic hypothesis does not explain why reliable matches are no longer observed once intramodal dimensions with congruent connotations are presented (e.g., visually presented shapes and colour do not appear to correspond).

**Keywords:** crossmodal correspondences, internet-based testing, intramodal correspondences, semantic hypothesis, semantic differential technique, sound symbolism.

## 1 Introduction

The last few years have seen a rapid growth of interest in the study of crossmodal correspondences (see Spence, 2011, 2012, for reviews). The term itself is one of many (see Spence, 2011, for various others) that have been used by researchers over the years in order to describe the fact that neurologically normal human observers appear to match objects, features, or dimensions of experience across sensory modalities, including in cases where they do not seem to be regularly co-experienced in the environment (see Deroy, Crisinel, & Spence, in press; Spence & Deroy, 2012). Initially, such intuitive crossmodal matches often strike us as surprising, not to say arbitrary. When, for example, people are asked to match non-words like "Takete" or "Kiki," and "Maluma" or "Bouba" to angular and rounded shapes (Köhler, 1929, 1947, Ramachandran & Hubbard, 2001), they respond consistently, matching "Takete" and "Kiki" to the angular shape, and "Maluma" and "Bouba" to the rounded one. When asked which colour is heavier, red or yellow, people consistently tend to say that red is the heavier of the two (Alexander & Shansky, 1976); they also intuitively associate brighter surfaces with higher pitched sounds, and darker ones with lower pitched sounds (Ludwig, Adachi, & Matzuzawa, 2011; see also O'Mahony, 1983).

One obvious limitation with many of these subjective questions, tested on only a limited number of participants, is that they do not reveal what drives/underlies the crossmodal matches. Over the years, several different explanations have been put forward, including the popular semantic hypothesis inspired by work on the semantic differential technique popularized by Osgood, Suci, and Tannenbaum (1957; see

also Osgood, 1960; Snider & Osgood, 1969). The idea here is that the presented objects are categorized along a certain number of common dimensions (such as active/passive; good/bad; dominant/submissive, etc.) and match if they fall in the same dimensional region as one another along some number of these dimensions. For instance, bright and high pitch are both "active," and therefore considered as more intuitively congruent than bright and low pitch. A similar hypothesis has also been adopted very recently by Walker and Walker (2012) and defended, albeit without commitment to specific conceptual dimensions, by Martino and Marks (1999, 2001; see also Palmer, Schloss, Xu, & Prado-León, 2013).

The lack of specificity, and the plurality of dimensions at stake, makes it feasible to criticize the semantic hypothesis as being possibly rather ad hoc: It is, after all, possible to find a common feature for any two objects one can think of, and then reconstruct it as *the* dimension that governed their matching. It is also possible that the participants in such studies may try to guess what sort of conceptual association the experimenter(s) had in mind when selecting a series of stimuli, making the effect an artifact of the forced-choice task more than the test of a pre-existing intuition (e.g., Pratt, 1990; Waterman, Blades, & Spencer, 2000). The goal of the present study was therefore to explore ways in which the semantic hypothesis could be put to the test empirically.

A first test here, inspired by the title of an article by Brown (1958), was to find some crossmodal matches which are semantically sound, and yet do not give rise to consistent intuitions. In the title of his review of Osgood et al.'s (1957) book on the semantic differential technique, Brown, for instance, considered the prosaic nature of a question like "*Is a boulder sweet or sour?*" Here, we tested the generality of this, and other crossmodal correspondences (such as that lemons are fast and that red is heavier than yellow), by utilizing an internet-based testing methodology which was trialed on the occasion of this study. We also tested a number of other similar associations involving non-relative or qualitative features such as "sweet–sour" that have frequently been endorsed in the literature (i.e., that lemons are fast and that red is heavier than yellow) to see whether the kind of effect obtained with boulders would generalize to these other cases. Similarly, we tested several other associations with selected features (sharp vs. round; red vs. yellow; rough vs. soft) that were already assigned a place on specific semantic dimensions by previous researchers/experiments to see whether or not they would generate intuitive matches that would be consistent with the semantic hypothesis.

A related prediction, which the defenders of the semantic hypothesis have not explicitly explored at this point, concerns intramodal correspondences: If all objects and features can be categorized according to certain common dimensions, and if any two features associated with the same pole on a given dimension should be matched by participants—this should happen regardless of whether those features are taken from the same versus different sensory modality. In this sense, one would expect people to match active "angular" to active "yellow" or "rough" within the visual modality, as much as they match active "up" and "high pitch" across modalities (e.g., across vision, audition, and touch; see Evans & Treisman, 2010; Occelli, Spence, & Zampini, 2009).

A final inference that can be drawn from the semantic hypothesis would appear to be that features that are easier to place in terms of certain crucial semantic dimensions (such as active/passive; good/bad; dominant/submissive) should lead to more frequent or confident matches by participants. For instance, one can think that the success of the Bouba/Kiki effect (e.g., Bremner et al., 2013; Köhler, 1929, 1947; Ramachandran & Hubbard, 2005) can be attributed to the fact that the sounds and/or the figures are easy to place on a scale that is anchored by the labels "passive" and "active" or "bad" and "good." Here, we tested whether participants would find this placing easier (i.e., more intuitive) by adding cues of the same polarity to the shape (such as adding cues like "happy" to the sharp shape, to make it more active, dominant, and good) or more difficult by adding cues going in the opposite direction (such as sad to the sharp shape, to make it *less* active, dominant, or good). In the present study, we tested these three hypotheses in a block of trials using a randomized internet testing method, first in an English-speaking environment that was highly controlled (Experiment 1), and then in a much broader selection of participants from different parts of the world in a somewhat less controlled manner (Experiment 2), in order to examine any differences that might be attributable to a participant's prior linguistic or cultural exposure (Eitan & Timmers, 2010). The consequences of the results obtained here for the semantic explanation of crossmodal matches are discussed below.

## 2 Experiment 1: Crossmodal matches and semantic dimensions

In Experiment 1, we tested three sets of predictions related to the semantic hypothesis of crossmodal matching. First, we were interested in determining whether certain matches that would challenge this

explanation also existed—primarily those involving qualitative (that is, metathetic, see Smith & Sera, 1992; Spence, 2011; Stevens, 1957) dimensions such as for hue or taste quality, or whole complex objects like boulders or fruits. Second, we wanted to test whether those features associated with the same polarity on specific dimensions would also be associated intramodally (e.g., shape and colours; colours and visually displayed textures). Third, we examined whether adding cues that made a certain polarity more salient on a particular semantically relevant dimension (e.g., happy for active sharpness) would lead to more confident or faster responding by participants on a classical crossmodal matching task ("Bouba–Kiki" being matched with round–angular shapes; Bremner et al., 2013; Köhler, 1929, 1947).

## 2.1 Methods

### 2.1.1 Participants

Eighty-one undergraduate students (71 female and 10 male) from York St. John University (UK) volunteered to take part in this study as part of a first year practical class. The participants ranged in age from 18 to 32 years (mean of 19.4 years). All of the participants provided informed consent prior to the study and the experimental protocol was approved by the University Ethics Committee. Participants reported their country of origin as United Kingdom (80) and Ireland (1). The average time taken to complete the study was 425 seconds (standard deviation 72 seconds).

### 2.1.2 Stimuli

Text descriptors and images were used as stimuli. The participants always had two response options (either text- or image-based) for each of the stimuli that were presented. The stimuli consisted of a subset presented in one configuration only ("boulder," "heavier," "prune," "yellow") and a subset of stimuli (blob, star, rectangle) that were varied parametrically across three factors (shape, colour, emotion; e.g., a blob that is red and has a smiley face). The stimuli and their possible response options available to participants are shown in Figure 1. We selected a variety of non-words that have been shown to correspond to "Bouba" (Bouba, Maluma, Lomoro, Mamima, Muromu, Malomu) and "Kiki" (Kiki, Takete, Kiriki, Teziki, Kichiki, Kekiti) in previous studies of sound symbolism (see Nielsen & Rendall, 2011). These non-words were chosen randomly from each list as needed (one randomly chosen word from each list was presented five times and the remaining words six times each). The stimuli were displayed against a grey background (RGB 217, 217, 217).

### 2.1.3 Apparatus

The participants completed the experiment in one of four lab classes on PC computers. The experiment was conducted on the internet through the Adobe Flash based Xperiment software (http://www.xperiment.mobi downloaded on 15/10/12). Testing was conducted in parallel on a number of computers in the psychology laboratory. The monitor resolution was set at 1,024 × 768 and had a screen size of 38.1 cm, corner to corner. All of the stimuli were clearly visible, each filling approximately 15% of the screen.

### 2.1.4 Design

A repeated measures design was used with all of the participants undertaking all of the experimental trials. The dependent variables were the response chosen (two possibilities), the reaction time (RT), and the confidence rating.

### 2.1.5 Procedure

The procedure was explained verbally to the participants by the experimenter and, at the start of the study, by means of an illustration (see Figure 2). The participants were instructed to follow the instructions that were presented on the screen, to remain quiet after completing the study, and to initiate the study when they were ready. All of the experimental trials followed the same procedure. A question mark was always presented in the centre of the screen for the duration of each trial. The test stimulus was positioned immediately above the question mark. The participant dragged the test stimulus to one of the two possible response options displayed equidistant to the left and right of the midline in the lower half of the screen. This response was recorded automatically. Left- and right-option positioning was randomized across trials and participants, as was the order in which the trials were presented. The time from the appearance of the stimulus up until when the stimulus was "dropped in place" over a

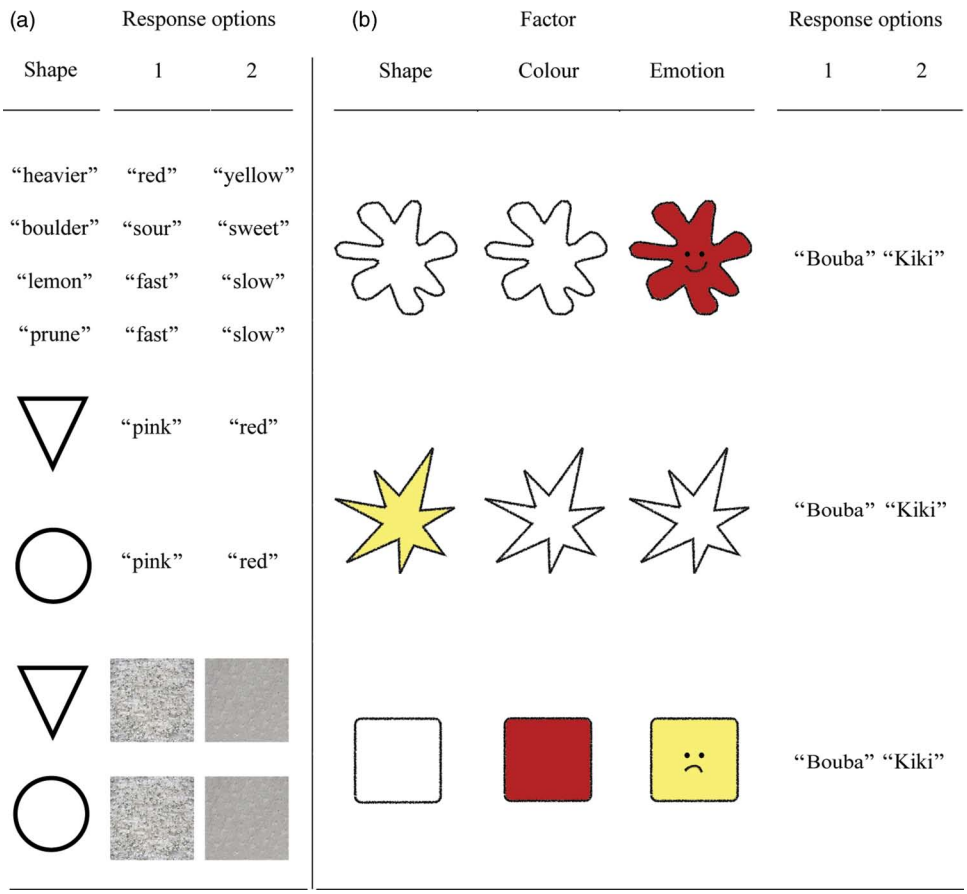


**Figure 1.** Stimuli used in Experiment 1 together with the two response options. (a) The subset of stimuli was presented in a single configuration only (no parametric variation); (b) the subset of stimuli that were parametrically varied in terms of their shape, colour, and emotion.

given response option by the participant was also recorded. Upon selecting a response, the participants were asked "How confident are you that other people will respond in the same way as you?" The participants made their responses on a five-point horizontal scale shown above the question mark with the following options, arranged from left to right: "very unconfident," "unconfident," "uncertain," "confident," "very confident." After responding, a "continue" button appeared. Pressing this button cleared the screen and initiated the next trial after a 500 ms pause. There were 35 trials per participant, which took approximately 10 minutes to complete. After the completion of the study, all of the participants were debriefed as to the nature of the study.

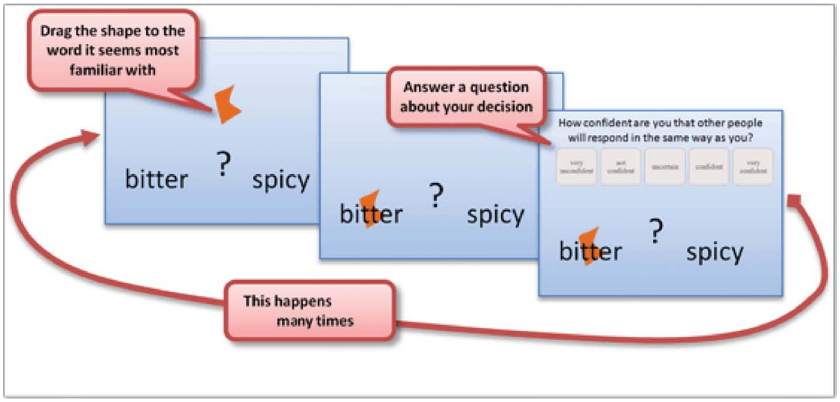


**Figure 2.** An illustration of the experimental procedure presented to the participants at the start of Experiment 1.

## 2.2 Results

The results (see Figure 3) clearly demonstrate that lemons and prunes differ in terms of their allocation to the labels "fast" and "slow," $\chi^2(1) = 40.55$, $p < 0.001$: Lemons were significantly more likely to be labelled as "fast" $\chi^2(1) = 9.00$, $p < 0.01$, whereas prunes were significantly more likely to be labelled as "slow" $\chi^2(1) = 34.68$, $p < 0.001$. In terms of the odds ratios, lemons were twice as likely to be labelled as "fast" than as "slow" (respective counts 54, 27), whilst prunes were 4.79 times more likely to be labelled as "slow" than as "fast" (67, 14). Boulders were 3.50 times more likely to be labelled as "sour" than as "sweet," $\chi^2(1) = 25.00$, $p < 0.001$ (63, 18). Meanwhile, "heavier" was 8.00 times more likely to be labelled as "red" than as "yellow" $\chi^2(1) = 49.00$, $p < 0.001$ (72, 9). There was no association between the colour descriptor chosen ("pink" or "red") for shape (circle, triangle), and the material-image (rough, smooth) chosen for shape.

Participants' confidence ratings were recoded numerically. A "very confident" response was assigned a value of 4, while a "very unconfident" response was assigned a value of 0; intermediate ratings were assigned appropriate intermediate values. In the subsequent analyses, confidence was used as the dependent variable. An ANOVA with shape (circle or triangle) as a repeated-measures factor and material-image (hard or soft) as a between-participants factor revealed a significant interaction term $F(1, 158) = 4.94$, $p < 0.05$, and post hoc LSD tests revealed that the circle was more confidently rated as "soft" than as "hard" ($p < 0.05$; respective means 2.61 and 2.14; in terms of count, the number of participants rating circles as "soft" was 38 and as "hard" was 43). There was also an interaction in an ANOVA with fruit (lemon and prune) and fastness ("slow" and "fast") as factors, $F(1, 158) = 9.29$, $p < 0.01$, with prunes being more confidently rated by participants as "slow" than as "fast" ($p < 0.05$), whilst lemons were more confidently rated as being "fast" than "slow" ($p < 0.05$). Finally, "heavier"

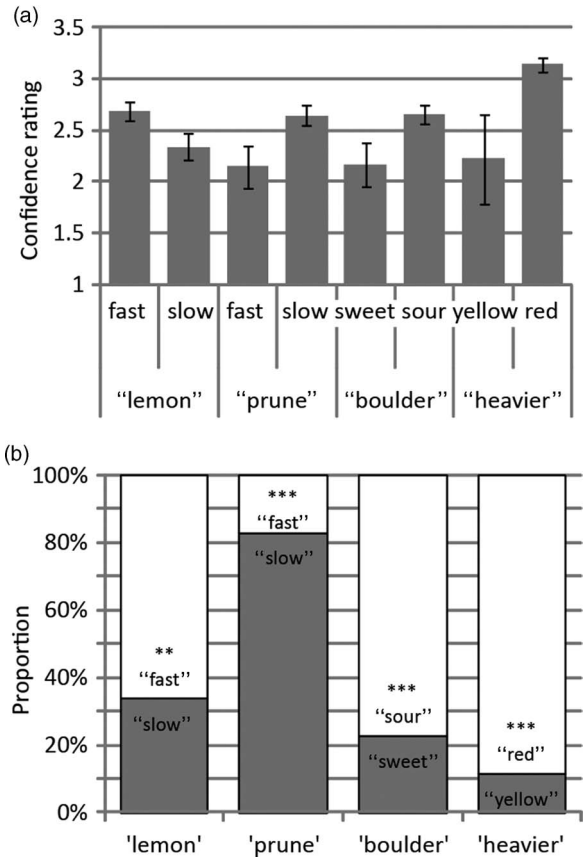


**Figure 3.** (a) Bar graph showing participants' confidence ratings when allocating stimuli to the different descriptors (error bars 2 SEM). (b) Bar graph showing proportion of different stimuli allocated to the various descriptors (**$p < 0.01$, ***$p < 0.001$, for participants favouring one response significantly over the other; according to $\chi^2$ tests).

was more confidently rated as "red" than as "yellow," $t(79) = 3.73$, $p < 0.001$. The boulder, likewise, was more confidently rated as "sour" than "sweet," $t(79) = 2.40$, $p < 0.05$ (respective means 2.65, 2.17).

A log-linear analysis was performed, using text ("Bouba" and "Kiki") × shape (blob, rectangle, star) × colour (red, white, yellow) × emotion (happy, neutral, sad) as the variables. The four-way log-linear analysis produced a final model that retained the text × shape × colour and text × emotion interactions. The likelihood ratio of this model was $\chi^2(32) = 14.83$, $p = 1.00$, indicating that it provided a good fit to the data. There was a trend effect for the text × shape × colour interaction, $\chi^2(2) = 8.99$, $p = 0.062$. The interactions between text and shape, $\chi^2(2) = 276.16$, $p < 0.001$, and between text and emotion, $\chi^2(2) = 24.09$, $p < 0\ .001$, were significant. In order to break the former interaction down, separate $\chi^2$ tests were performed with each type of shape to determine whether they were assigned as "Bouba" or "Kiki" (Text). Blob $\chi^2(1) = 67.00$, $p < 0.001$, rectangles, $\chi^2(1) = 61.07$, $p < 0.001$, and star, $\chi^2(1) = 143.11$, $p < 0.001$ were all allocated more to one text descriptor than to the other (see Figure 4a). Odds ratios indicated that the rectangles and blob shapes were 1.81 and 1.87 times more likely to be rated as "Bouba" than as "Kiki," respectively. By contrast, the star was 2.59 times more likely to be rated as "Kiki" (than as "Bouba"), explaining the original interaction and implying a genuine difference in the descriptor chosen between the star and the other two shapes. No lower-order significant factors which interacted significantly in higher order effects are reported.

The text by emotion interaction was broken down in the same fashion as the previous text and shape interaction in order to determine whether there were any significant differences in the assignment of "Bouba" or "Kiki." This was only true for sad stimuli, $\chi^2(1) = 24.27$, $p < 0.001$, which were 1.45 times more likely to be rated as "Bouba" than as "Kiki" (see Figure 5a).

A four-way between-participants ANOVA was conducted on the confidence rating data (note that one data point was missing from this analysis). The factors consisted of the same independent variables as above. The interaction between text selected and shape was significant, $F(2, 2{,}132) = 21.18$, $p < 0.001$ (see Figure 4b). Post hoc LSD tests revealed that the participants were more confident in assigning the label "Kiki" than the label "Bouba" to the stars ($p < 0.001$), and, conversely, the label "Bouba" (rather than "Kiki") to the blobs ($p < 0.001$) and rectangles ($p < 0.01$).

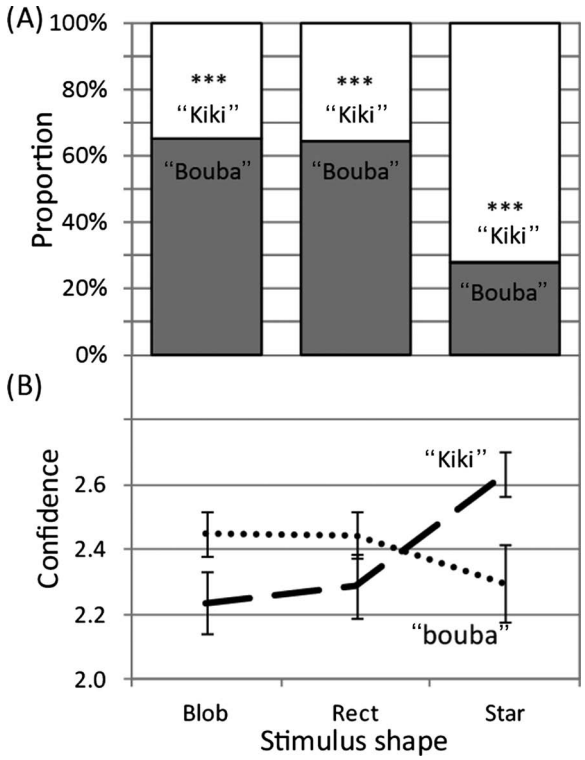


**Figure 4.** (a) Bar graph showing the proportion of blob, rectangle, and star stimuli being associated with the words "Bouba" and "Kiki" (grey and white bars, respectively, where ***$p < 0.001$) in Experiment 1; (b) line graph showing "Bouba" and "Kiki" confidence ratings for the blob, rectangle, and star stimuli (error bars 2 SEM).

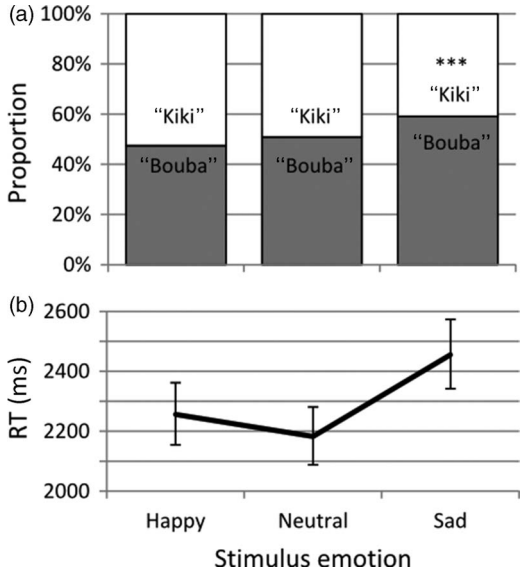


**Figure 5.** (a) Bar graph showing the proportion of happy, neutral, and sad stimuli associated with the words "Bouba" and "Kiki" in Experiment 1 (where ***$p < 0.001$). (b) Line graph showing RT (in ms) to assign happy, neutral, and sad stimuli as either "Bouba" or "Kiki"; error bars (2 SEM) were calculated for log RT but then reverted to normal scale via an inverse log function for ease of presentation.

The RT data were leptokurtic and positively skewed, and hence, in order to help correct for this, were log-transformed (see Field, 2005). Outlying data for RTs to stimuli reported as "Bouba" and those reported as "Kiki" were screened separately and corrected for (<1% in both cases; outliers are henceforth defined as exceeding ±3 standard deviations surrounding the mean, and henceforth corrected for by being replaced with the next most extreme but non-outlying data point). For "single configuration stimuli" (Figure 1a), two two-way ANOVAs were conducted with log RT as the dependent variable, a repeated-measures factor of stimulus (circle or triangle), and also a between-participants response factor of either pink/red or hard-/soft-material (see the bottom four rows in Figure 1a). Four independent sample *t*-tests were also conducted in order to see whether stimulus (either "heavier," "boulder," "lemon," "prune") were allocated more so to one response or to another (responses were respectively "red"/"yellow," "sweet"/"sour," "fast"/"slow" and "fast"/"slow" again; see the top four rows in Figure 1a). There were no significant effects on log RT between allocations in those trials with a circle, triangle, "heavier," and "boulder" as the stimuli. There was a trend towards an effect for the "lemon" and "prune" $F(1, 158) = 3.52$, $p = 0.063$, with participants taking more time to assign "lemons" than to assign "prunes" (computed averages 3,038 and 2,529 ms, respectively). For "parametrically varied" stimuli (Figure 1b), a four-way between-participants ANOVA was conducted with log RT as the dependent variable and shape, colour and emotion as independent variables. The only significant factor to emerge from this analysis was emotion, $F(2, 2{,}133) = 5.24$, $p < 0.01$ (see Figure 5b). Post hoc LSD tests revealed that participants responded more slowly to sad stimuli than to either happy ($p < 0.01$) or neutral ($p < 0.001$) stimuli (average computed RTs were 2,455, 2,256 and 2,182 ms, respectively). This finding has been reported before (one explanation is that negative stimuli might be processed more slowly than positive and neutral stimuli to avoid potentially dangerous errors, see Ihssen & Keil, 2013).

## 2.3 Discussion

The results of Experiment 1 can be divided into two parts: The re-testing of already documented crossmodal matches, and the testing of new correspondences. Our results establish that boulders are associated with sourness while prunes are associated with slowness. By contrast, no significant matching pattern was found intramodally, that is, between shapes (circle/triangle) and texture (soft/round), or shapes and colour (red/pink)—although those participants who rated circles as soft were more confident of their choice than those who rated them as rough. Regarding those crossmodal matches that have already been documented in the literature, our results are consistent with previous or expected

results (such as the fastness of lemons, see Smith, 2012, and the heaviness of red or the attribution of "Kiki" to star-shaped figures). That said, our results also demonstrate that intermediate figures (rectangle with rounded corners) are rated as Bouba, and that, in some circumstances, emotional cues can override the intuitive associations that exist between shapes and sounds, as sad faces embedded in either of the shapes have a tendency to be associated with "Bouba." This result can be explained in terms of the semantic hypothesis by asserting that the sound "Bouba" is associated with passive stimuli and that sad faces are also more strongly associated with passivity than are shapes. However, the same result underscores the risk of the ad hoc application of the semantic hypothesis: if "Bouba" is associated with the blobby shape, and the blobby shape can be described as good (or at least as less bad) than the spiky star shape, then it is rather surprising that "Bouba" is also associated with sad, i.e., negative emotions. No effect on the "Bouba"–"Kiki" matching was documented in the case of colour (red/yellow), also creating something of a tension for the semantic hypothesis: If the semantic dimensions attributed to "Kiki" are active and/or bad (explaining the matching with angularity), then one would expect a match with "red" which is often recognized as alerting or enlivening (Elliot & Maier, 2007; Hogg, 1969).

## 3 Experiment 2: Crossmodal matches across linguistic borders

In a follow-up study, we tested whether the previous results could be attributed to cross-culturally semantic dimensions that are common to all objects (as posited by the semantic hypothesis) and whether they were universal as often assumed in the field of crossmodal correspondences (see Bremner et al., 2013, for a discussion).

### 3.1 Methods

#### *3.1.1 Participants*

Eighty-two individuals (39 female and 43 male) recruited from Amazon's Mechanical Turk took part in this study in exchange for a payment of 0.80 US dollars on the April 19, 2013, from 12:00 GMT onwards over a four-hour period (see Crump, McDonnell, & Gureckis, 2013, for a methodological overview). The participants ranged in age from 22 to 68 years (mean of 35.3 years). All of the participants provided informed consent prior to their taking part in the study. The participants' countries of origin are shown in Figure 6 (derived from the IP address via the GeoLite City database; downloaded from http://dev.maxmind.com/geoip/legacy/geolite on May 1, 2013). The average time taken to complete the study was 240 seconds (standard deviation 78 seconds).

#### *3.1.2 Stimuli, apparatus, design, and procedure*

This experiment was almost identical to the previous one in terms of the stimuli, design, and procedure, but excluded verbal instruction of the procedure to the participant by the experimenter (given the remote locations in which the participants completed the study). The apparatus varied across participants. Two participants performed the study on Apple computers (version 10.7.5 and 10.6.8) and the remainder on machines running Windows (20 participants on Windows XP, 44 on Windows 7 and 10 on Windows 8). The most common resolution of the participants' monitors was 1,366 × 768 (28 participants) and the mean resolution was 1,422 × 867 (their respective standard deviations were 243 and 136).


**Figure 6.** World map showing the geo-location of participants' data ©2012 Google, INEGI, MapLink (red circle = location of a female participant; blue circle = male participant).

## 3.2 Results

The same statistical procedures reported in Experiment 1 were also used here. To summarize, there was little difference between the results of the two experiments. Once again, the results reported here demonstrate that lemons and prunes differed in terms of the allocation to the labels "fast" and "slow," $\chi^2(1) = 11.83$, $p < 0.001$: lemons were significantly more likely to be labelled as "fast" $\chi^2(1) = 3.95$, $p = 0.05$, whereas prunes were significantly more likely to be labelled as "slow" $\chi^2(1) = 8.24$, $p < 0.01$. In terms of the odds ratios (see Figure 7), lemons were 1.56 times more likely to be labelled as "fast" than as "slow" (respective counts 50, 32), whilst prunes were 1.93 times more likely to be labelled as "slow" than as "fast" (54, 28). Boulders were 1.73 times more likely to be labelled as "sour" than as "sweet," $\chi^2(1) = 5.90$, $p < 0.02$ (52, 30). Meanwhile, "heavier" was 4.13 times more likely to be labelled with the "red" descriptor than with the "yellow" descriptor, $\chi^2(1) = 30.49$, $p < 0.001$ (66, 16). In contrast to the results of Experiment 1, there was an association between the colour descriptor chosen ("pink" or "red") and shape, with circles being 1.93 times more likely to be labelled as red than as pink $\chi^2(1) = 8.24$, $p < 0.01$ (54, 28). There was no association between the triangle and colour, and as with Experiment 1, no association between material-image (rough and smooth) and shape. In terms of the confidence analyses, there was only a significant effect of text for boulders $F(1, 80) = 6.48$, $p < 0.02$, with stimuli rated as "sour" being more confidently rated than were the "sweet" stimuli (means were 2.56 and 2.13, respectively). In terms of log RTs (as before the data were corrected via a log transform due to their being leptokurtic and positively skewed), the only significant effect was with the two-way ANOVA with text ("red" or "pink") and stimulus (triangle, circle); there was an interaction between the factors $F(1, 160) = 4.01$, $p < 0.05$, with triangles being rated more quickly as red than as pink ($p < 0.05$; means of 1,763 ms and 2,403 ms, respectively).

The log-linear analysis reported in Experiment 1 was also performed here and produced similar results. The likelihood ratio of the model was $\chi^2(44) = 14.24$, $p = 1.00$, and it retained both interactions of text × shape $\chi^2(2) = 48.31$, $p < 0.001$ and text × emotion $\chi^2(2) = 15.56$, $p < 0.001$. In order to break the former interaction down, separate $\chi^2$ tests were performed with each type of shape in order to determine whether they were assigned to "Bouba" or to "Kiki" (text). Rectangle, $\chi^2(1) = 5.90$, $p < 0.05$ and star, $\chi^2(1) = 40.09$, $p < 0.001$, were allocated more to one text descriptor than to the other (see Figure 8a); blob, $\chi^2(1) \times 3.122$, $p < 0.077$, did so at trend level. Odds ratios indicated that the rectangle and blob shapes were 1.20 and 1.14 times more likely to be rated as "Bouba" than as "Kiki," respectively. By contrast, the star was 1.61 times more likely to be rated as "Kiki" (than as "Bouba"), thus explaining the original interaction, and implying a genuine difference in chosen descriptor between the star and the other two shapes.

The interaction between text and emotion was broken down in the same fashion, in order to determine whether there were any significant differences in the assignment of "Bouba" or "Kiki." Happy stimuli $\chi^2(1) = 7.42$, $p < 0.01$ were 1.22 times more likely to be rated as "Kiki" than as "Bouba," whilst an opposite trend was observed for the sad stimuli, $\chi^2(1) = 5.9$, $p < 0.02$, which were 1.20 times more likely to be rated as "Bouba" than as "Kiki" (see Figure 9). The neutral stimuli were 1.14

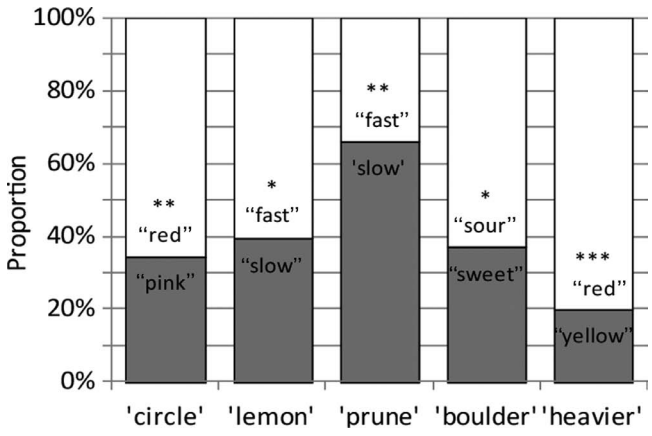


**Figure 7.** Bar graph showing the proportion of different stimuli to descriptors in Experiment 2 (where *$p < 0.05$; **$p < 0.01$; and ***$p < 0.001$).

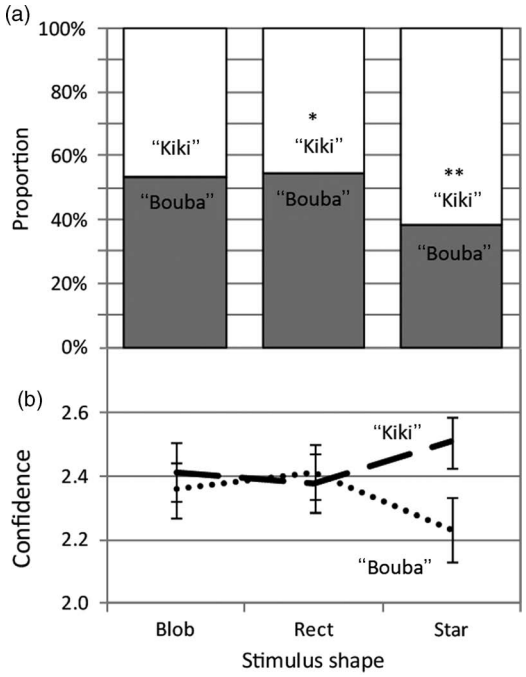


**Figure 8.** (a) Bar graph highlighting the proportion of blob, rectangle, and star stimuli allocated to the words "Bouba" and "Kiki" (grey and white bars, respectively) in Experiment 2 (where *$p < 0.05$ and **$p < 0.01$); (b) line graph showing "Bouba" and "Kiki" confidence ratings for blob, rectangle, and star stimuli (error bars 2 SEM).

times more likely to be rated as "Kiki" than as "Bouba," although this trend just failed to reach statistical significance $\chi^2(1) = 3.39$, $p = 0.066$.

A four-way between-participants ANOVA was conducted on the confidence rating data. The factors consisted of the same independent variables as above. Although the four-way interaction was significant, $F(8, 2{,}160) = 3.01$[1] $p < 0.01$, the exploration of this interaction will not be reported here as it does not aid in the testing of our hypotheses. The interaction between shape and text was significant, $F(2, 2{,}160) = 6.13$, $p \leq 0.005$ (see Figure 8b). Post hoc LSD tests revealed that the participants were more confident in assigning the label "Kiki" than the label "Bouba" to the stars ($p < 0.001$). There was, however, no difference in terms of our participants' confidence when it came to assigning blobs or rectangles to the "Kiki" and "Bouba" responses, respectively.

A four-way between-participants ANOVA was conducted with log RT as the dependent variable and with the same independent variables as described in Experiment 1 (outlying data for RTs to stimuli reported as "Bouba" and those reported as "Kiki" were screened separately and corrected for; <2% in both cases). The only significant factor to emerge from this analysis was shape, $F(2, 2{,}160) = 3.02$, $p < 0.05$. Post hoc LSD tests revealed that participants responded more quickly to blob stimuli than to

[1]The appropriate interpretation of this four-way interaction was determined by running a number of separate ANOVA for each shape with the remaining factors. For the star, only the main effect of text was significant, $F(1, 720) = 18.26$, $p < 0.001$), with stimuli rated more confidently as "Kiki" than "Bouba" (means of 2.51 versus 2.23, respectively). For the blobby shape, the three-way interaction between colour, emotion, and text was significant $F(4, 720) = 4.30$, $p < 0.01$. Separate ANOVAs were conducted for each blob colour (collapsing over emotion), and for each blob emotion (collapsing over colour); there were no significant effects from the blob emotion ANOVAs. In the blob colour ANOVAs, however, the emotion × text interaction was significant for both white and yellow blobs, $F(2, 240) = 6.10$, $p < 0.01$, and $F(2, 240) = 3.44$, $p < 0.05$. For white blobs (happy, neutral and sad mean ratings for white "Bouba" were 2.11, 2.71, 2.31, respectively, and for white "Kiki" were 2.43, 2.11, 2.35), neutral emotion "Bouba" confidence ratings were both larger than neutral "Kiki" ratings ($p < 0.01$) as well as happy ($p < 0.01$) and sad ($p < 0.05$) "Bouba". For yellow blobs ratings (happy, neutral and sad mean ratings for yellow "Bouba" were 2.45, 2.18, 2.32, respectively, and for yellow 'Kiki' 2.63, 2.68, 2.13), neutral emotion "Kiki" confidence ratings were both larger than "Bouba" ratings for neutral emotion stimuli ($p < 0.01$) as well as sad "Kiki" and "Bouba" ratings ($p < 0.001$, $p = 0.05$, respectively). Yellow neutral emotion "Bouba" ratings were rated lower than happy "Kiki" ratings ($p < 0.02$). Finally, happy "Kiki" ratings were higher than sad "Kiki" ratings ($p < 0.02$). There were no significant effects for red blobs.

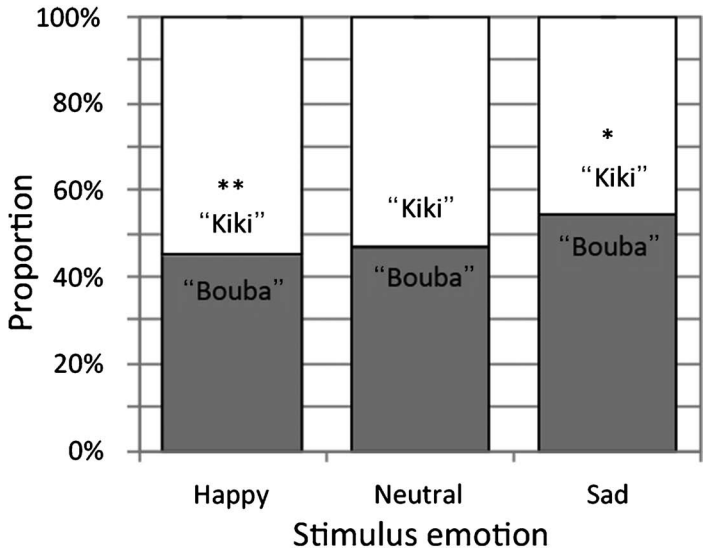


**Figure 9.** Bar graph showing the proportion of happy, neutral, and sad stimuli associated with the words "Bouba" and "Kiki" in Experiment 2 (where $*p < 0.05$ and $**p < 0.01$).

either rectangle ($p < 0.05$) or star ($p < 0.02$) stimuli (average computed RTs were 2,173, 2,345, and 2375 ms, respectively).

### 3.3 Discussion

The results of Experiment 2 confirm the cross-cultural existence of the crossmodal matchings that were originally documented in Experiment 1, although the cross-cultural aspect is tempered by the high proportion of North American participants relative to those from other locations. This result therefore confirms that people really do tend to associate boulders with "sour," "heavy" and "red," and converges on the conclusion that lemons are fast while prunes are slow. The distribution of responses was also similar, with convergence being more important for the heaviness of the colour red and the slowness of prunes than for the sourness of boulders and the speed of lemons. However, we also observed an intramodal correspondence between circles and red, which raises questions about the absence of intramodal matching in the previous set of participants. The results on the "Bouba"–"Kiki" matchings were also broadly similar to those obtained in Experiment 1 (and which have already been documented recently in a remote culture; see Bremner et al., 2013, for a cross-cultural study of the "Bouba"–"Kiki" effect), while showing some minor differences regarding the effects of emotional cues (here, both sad and happy faces were more often associated with "Kiki" than "Bouba"), thus suggesting that the effect of emotion should be more thoroughly investigated in future research.

## 4 General discussion

The results of the two experiments reported in the present study demonstrate the potential power of internet-based testing methods for those wanting to assess a variety of different crossmodal correspondences. Despite participants completing Experiment 2 unsupervised on a wide variety of hardware in various testing environments, the results of both experiments were satisfyingly congruent with one another (as also reported by Crump et al., 2013; Germine et al., 2012). The speed of data collection in Experiment 2 was in addition remarkably fast (82 participants in 4 hours) and very cost effective, and indeed perhaps "revolutionary" (Crump et al., 2013). Conceivably, the limiting factors for online research are now the limited number of sensory modalities that can be stimulated and the restricted range of sensors that can be used to collect data—smartphones do have an edge over computers in this regard, indicating the direction of the potential next "revolutionary" step in data collection (Miller, 2012). Our results also confirm the fact that, as intuited by Peter Walker (and others, Deroy, 2011; Smith, 2012), most people do indeed consider lemons to be fast. One of the novel findings to emerge from the results of the present study was that most people also associate prunes with slowness, an association that, as far as we are aware, has not been documented previously. The experimental technique used here allowed us to distinguish between the inter-individual (overall likelihood of matching

across all trials) and the intra-individual (confidence ratings plus RT) intuitive appeal of crossmodal correspondences, and to check the cross-cultural validity of the results.

According to the semantic hypothesis, surprising associations between sensory features or dimensions arise because the information available to different modalities is recoded into a more abstract, semantic format. The cross-cultural similarities are, in themselves, important to question the link between this "semantic" recoding and linguistic representations (as suggested for instance by Martino & Marks, 1999): In the case of thickness and pitch, the association seems to exist—at least, to introduce an interference in a perceptual task—only for Farsi participants who describe pitches as thin (na-zok) or thick (koloft) (see Dolscheid, Shayan, Majid, & Casasanto, 2013). Here, we found that intuitive crossmodal associations were widespread across cultural and linguistic groups.

Instead of language, it has been suggested that crossmodal associations are governed by general dimensions shared or "connoted" by the perceptual concepts applied to the stimuli (Walker & Walker, 2012). According to the semantic differential technique, as introduced by Osgood et al. (1957; Snider & Osgood, 1969), more than half a century ago, many concepts can be analysed (and then matched) in terms of dimensions anchored by pairs of polar adjectives; with the most commonly retained dimensions being active–passive, good–bad, and dominant–submissive (e.g., Proctor & Cho, 2006). According to this approach, for instance, angularity and bitterness are matched because both angular and bitter fall on the same "bad" end of the bad/good dimension (i.e., they are both potentially dangerous kinds of stimuli; Bar, 2007, 2011; Bar & Neta, 2006).

In answer to the question posed in the title of Brown's (1958) early paper (and reprinted in Snider & Osgood, 1969), there really is a crossmodal association (or correspondence) between boulders and sourness—ironically one for which there is even more inter-individual agreement than the successful "fast lemons" originally proposed by Peter Walker. That said, these crossmodal associations are weaker than the weight of hues, as assessed here (and borrowed from the research of Alexander & Shansky, 1976). The existence of a correspondence between boulders and sourness is at first quite supportive of the semantic hypothesis, as it seems not to fail on this example. The fact is that we presumably cannot rely on an individual's intuitions as necessarily reflecting a consensus (as Koriat, 2008, suggested for sound–symbolic associations between foreign words and their referents). Instead, we need to validate each intuition with experimentation, and this is where a fast, non-expensive and far-reaching new technique like the internet-based testing shows its strength.

Still, the existence of an intuitive matching between boulders and sourness reinforces a pre-existing challenge for the hypothesis, as one needs to explain how it is that qualitative, metathetic dimensions (e.g., of hue and taste quality; Spence, 2011) or natural objects with variable features (boulders) are assigned a specific polarity: Why, for example, should yellow, lemons and boulders all be more active?

We also examined a rarely discussed aspect of correspondences, that is, the presence of intramodal correspondences. The results reported here suggest that there may not be a systematic intramodal correspondence between shape and texture (soft/rough) although one would expect that both round and soft, and triangle and rough would be given the same polarity regarding certain semantic dimensions ("good" for the former, and "active" for the latter). There was also no systematic correspondence between colour and shape (except for the matching of roundness with redness in Experiment 2). This finding challenges Kandinsky's (1925) early claim that many instances of such matches exist.

In testing the "Bouba"/"Kiki" effect, we were also able to confirm that making the angularity/roundness contrast more salient (by using a star with many angles and a square with fewer) affected the intra-individual confidence of participants in the matching. Interestingly, the prediction generated by a consideration of the semantic differential technique (Osgood et al., 1957), that adding cues of sadness to the shape-cues of passivity or cues of happiness to shape-cues of activity would influence the matching between sounds and visual images, was not observed. What is more, the effect of emotional faces on matching also differed as a function of the regions of the world in which the participants were located, but were strong enough to substantially influence the performance of participants on the matching task in Experiment 1, with a sad schematic line drawing of a face overriding shape cues in the association with "Bouba." This result is, however, insufficient to refute the explanations of crossmodal correspondences in terms of the semantic differential technique—as it is possible to construct an explanation in terms of another shared dimension (say, for example, that both "Bouba " and "sad face" are passive). However, this result certainly points to the need for those who support the explanatory power of the semantic hypothesis to state precisely which dimension prevails and why, in a given situation.

More fundamentally, these results show the need to question whether the dimensions singled out by Osgood et al. (1957) should be the ones (or the only ones) used when explaining crossmodal correspondences. In this sense, Martino and Marks's (2001, p. 64) idea of an *"abstract semantic network that captures synesthetic correspondences"* more carefully avoids mentioning the dimensions inherited from work on the semantic differential technique, but could still be criticized for failing to generate precise refutable hypotheses (Platt, 1964).

## 5 Conclusions

Over the years, many researchers have been eager to explain surprising crossmodal matches in terms of a form of semantic hypothesis (e.g., Martino & Marks, 1999, 2001; Walker & Walker, 2012; see also Proctor & Cho, 2006): The idea is that people match sensory stimuli across different sensory modalities because they connote the same pole on a set of fundamental dimensions, such as active/passive or good/bad. Evaluating this hypothesis empirically is difficult and will certainly require examples used in its defence or against it to be robustly tested. Evaluating the hypothesis will also require researchers to test a number of predictions that follow on from the hypothesis. Here, we have pointed to four difficulties: The semantic hypothesis does not explain differences in frequency—or individual confidence—which exist between crossmodal associations built on the same model (fast lemons, slow prunes, sour boulders, and heavy red); it does not explain why certain qualitative dimensions get assigned to a specific polarity in the scaling; it only minimally predicts what happens when the connotations that are supposed to drive the associations are made more salient; finally, it does not explain why the matches are no longer observed once intramodal dimensions with congruent connotations are presented. Here, we would like to suggest that this does not mean that the semantic hypothesis does not account for parts of the reflective processes by which people rationalize their intuitive sense of crossmodal congruency, but that it lowers its ability to explain what determines these intuitions in the first place. In this respect, the use of internet-based testing protocols, which is now starting to become more widely used across various areas of psychological research (see Germine et al., 2012, for a recent review) will turn out to be crucial in reaching out to a more varied array of potential participants than those most commonly used in psychological experiments currently (i.e., those coming from Western, Educated, Industrialized, Rich, and Democratic societies—or WEIRD, see Henrich, Heine, & Norenzayan, 2010).

## References


Alexander, K. R., & Shansky, M. S. (1976). Influence of hue, value, and chroma on the perceived heaviness of colors. *Perception and Psychophysics, 19,* 72–74. doi:10.3758/BF03199388

Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences, 11*, 280–289. doi:10.1016/j.tics.2007.05.005

Bar, M. (2011). *Predictions in the brain: Using our past to generate a future*. Oxford: Oxford University Press.

Bar, M., & Neta, M. (2006). Humans prefer curved visual objects. *Psychological Science, 17,* 645–648. doi:10.1111/j.1467-9280.2006.01759.x

Bremner, A., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K., & Spence, C. (2013). Bouba and Kiki in Namibia? Western shape-symbolism does not extend to taste in a remote population. *Cognition, 126,* 165–172. doi:10.1016/j.cognition.2012.09.007

Brown, R. W. (1958). Is a boulder sweet or sour? *Contemporary Psychology, 3,* 113–115.

Crump, J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioural research. *PloS One, 8,* e57410. doi:10.1371/journal.pone.0057410

Deroy, O. (2011). Fast lemons and intuitive beliefs. International Cognition and Culture Institute. http://www.cognitionandculture.net/home/blog/13-ophelias-blog/804-fast-lemons-and-intuitive-beliefs.

Deroy, O., Crisinel, A.-S., & Spence, C. (2013). Crossmodal correspondences between odours and contingent features: Odours, musical notes, and arbitrary shapes. *Psychonomic Bulletin and Review*. advance online publication. doi:10.3758/s13423-013-0397-0

Dolscheid, S., Shayan, S., Majid, A., & Casasanto, D. (2013). The thickness of musical pitch: Psychophysical evidence for linguistic relativity. *Psychological Science, 24,* 613–621.doi:10.1177/0956797612457374

Eitan, Z., & Timmers, R. (2010). Beethoven's last piano sonata and those who follow crocodiles: Cross-domain mappings of auditory pitch in a musical context. *Cognition, 114,* 405–422. doi:10.1016/j.cognition.2009.10.013

Elliot, A., & Maier, M. (2007). Colour and psychological functioning. *Current Directions in Psychological Science, 16,* 250–254. doi:10.1111/j.1467-8721.2007.00514.x

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision, 10,* 1–12. doi:10.1167/10.1.6

Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). London: Sage Publications.

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin and Review, 19,* 847–857. doi:10.3758/s13423-012-0296-9

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33,* 61-135. doi:10.1017/S0140525X0999152X

Hogg, J. (1969). A principal components analysis of semantic differential judgements of single colors and color pairs. *Journal of General Psychology, 80,* 129–140. doi:10.1080/00221309.1969.9711279

Ihssen, N., & Keil, A. (2013). Accelerative and decelerative effects of hedonic valence and emotion arousal during visual scene processing. *Quarterly Journal of Experimental Psychology, 60,* 1276–1301. doi:10.1080/17470218.2012.737003

Kandinsky, W. (1925). *Point and line to plane*. New York: Dover Publications.

Köhler, W. (1929). *Gestalt psychology*. New York: Liveright.

Köhler, W. (1947). *Gestalt psychology: An introduction to new concepts in modern psychology*. New York: Liveright Publication.

Koriat, A. (2008). Subjective confidence in one's answers: The consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 945–959. doi:10.1037/0278-7393.34.4.945

Ludwig, V. U., Adachi, I., & Matzuzawa, T. (2011). Visuoauditory mappings between high luminance and high pitch are shared by chimpanzees (Pan troglodytes) and humans. *Proceedings of the National Academy of Sciences USA, 108,* 20661–20665. doi:10.1073/pnas.1112605108

Martino, G., & Marks, L. E. (1999). Perceptual and linguistic interactions in speeded classification: Tests of the semantic coding hypothesis. *Perception, 28,* 903–923. doi:10.1068/p2866

Martino, G., & Marks, L. E. (2001). Synesthesia: Strong and weak. *Current Directions in Psychological Science, 10,* 61–65. doi:10.1111/1467-8721.00116

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science, 7,* 221–237. doi:10.1177/1745691612441215

Nielsen, A., & Rendall, D. (2011). The sound of round: Evaluating the sound-symbolic role of consonants in the classic Takete-Maluma phenomenon. *Canadian Journal of Experimental Psychology, 65,* 115–124. doi:10.1037/a0022268

Occelli, V., Spence, C., & Zampini, M. (2009). Compatibility effects between sound frequency and tactile elevation. *Neuroreport, 20,* 793–797. doi:10.1097/WNR.0b013e32832b8069

O'Mahony, M. (1983). Gustatory responses to nongustatory stimuli. *Perception, 12,* 627–633. doi:10.1068/p120627

Osgood, C. E. (1960). The cross-cultural generality of visual–verbal synesthetic tendencies. *Behavioral Science, 5,* 146–169. doi:10.1002/bs.3830050204

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Palmer, S. E., Schloss, K. B., Xu, Z. X., & Prado-León, L. (2013). Color, music, and emotion. *Proceedings of the National Academy of Sciences, 110,* 8836–8841. doi:10.1073/pnas.1212562110

Platt, J. R. (1964). Strong inference. *Science, 146,* 347–353. doi:10.1126/science.146.3642.347

Pratt, C. (1990). On asking children – and adults – bizarre questions. *First Language, 10,* 167–175. doi:10.1177/014272379001002905

Proctor, R. W., & Cho, Y. S. (2006). Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin, 132,* 416–442. doi:10.1037/0033-2909.132.3.416

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia: A window into perception, thought and language. *Journal of Consciousness Studies, 8*, 3–34.

Ramachandran, V. S., & Hubbard, E. M. (2005). The emergence of the human mind: Some clues from synesthesia. In L. C. Robertson, & N. Sagiv (Eds.), *Synesthesia: Perspectives from cognitive neuroscience* (pp. 147–192). Oxford: Oxford University Press.

Smith, B. (2012). Lemons are fast: Cross-sensory correspondences and naïve conceptions of natural phenomena. The Edge, EDGE Annual Question 2012: http://edge.org/annual-question, at http://www.edge.org/responses/what-is-your-favorite-deep-elegant-or-beautiful-explanation

Smith, L. B., & Sera, M. D. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology, 24,* 99–142. doi:10.1016/0010-0285(92)90004-L

Snider, J. G., & Osgood, C. E. (1969). *Semantic differential technique: A sourcebook.* Chicago: Aldine Publishing.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, and Psychophysics, 73,* 971–995. doi:10.3758/s13414-010-0073-7

Spence, C. (2012). Managing sensory expectations concerning products and brands: Capitalizing on the potential of sound and shape symbolism. *Journal of Consumer Psychology, 22,* 37–54. doi:10.1016/j.jcps.2011.09.004

Spence, C., & Deroy, O. (2012). Are chimpanzees really synaesthetic? *i-Perception, 3,* 316–318. doi:10.1068/i0526ic

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64,* 153–181. doi:10.1037/h0046162

Walker, P., & Walker, L. (2012). Size-brightness correspondence: Crosstalk and congruity among dimensions of connotative meaning. *Attention, Perception, and Psychophysics, 74,* 1226–1240. doi:10.3758/s13414-012-0297-9

Waterman, A. H., Blades, M., & Spencer, C. P. (2000). Do children try to answer nonsensical questions? *British Journal of Developmental Psychology, 18,* 211–226. doi:10.1348/026151000165652