Social Cognition, Vol. 29, No. 5, 2011, pp. 547–576

CATEGORY BOUNDARIES AND CATEGORY LABELS: WHEN DOES A CATEGORY NAME INFLUENCE THE PERCEIVED SIMILARITY OF CATEGORY MEMBERS?

Francesco Foroni
Utrecht University

Myron Rothbart
University of Oregon

Three experiments examined the effect of verbal labels on the perception of category members. Participants were presented with silhouette drawings of female body types, ordered on a continuum from very thin to very heavy, and asked to judge the degree of similarity between pairs, as well as absolute weight of each silhouette. The presence/absence of category boundaries and labels were experimentally manipulated (Exp. 1-3), as was the "strength" of the labels (Exp. 2 and 3), their source (Exp. 1 and 2), and their implications (Exp. 3). The presence of a label, even when self-generated, showed clear effects on judgment: labels consistently increased within-category similarity (assimilation), and reduced across-category similarity (contrast). The judged strength of the verbal labels was correlated with the strength of categorization effects.

There is an old joke about a peasant living on the border of Poland and Russia, distressed at not knowing the exact location of his farm. When surveyors, hired at great expense, finally determined the farm to be barely inside the Polish border, the farmer exclaimed with great relief: "Thank God! No more Russian winters!" There are many levels of irony in this story, including the belief that a country's climate ends abruptly at its national boundaries. Stated differently, objectively tiny differences between points on a continuum are perceived as large, or discontinuous, due to the interposition of a category boundary, even when the boundary placement itself is arbitrary. This mode of thinking is not limited to the poorly educated:

This research was supported by National Institute of Mental Health Grant MH40662 to the second author. We wish to thank Mary Rothbart, Bernardette Park, Ellen Peters, and Gale Pearce for their help and suggestions with this project and with an early draft of the article.

Correspondence concerning this article should be addressed to Francesco Foroni, Faculty of Social & Behavioral Sciences, University of Utrecht, Utrecht - The Netherlands. E-mail: f.foroni@uu.nl.

Krueger & Clement (1994) found that college students' estimates of daily temperatures were assimilated toward monthly averages, leading to an overestimation of temperature differences for days on opposite sides of monthly boundaries.

More generally, imposing arbitrary category boundaries on a meaningful continuum provides a useful paradigm for understanding the effects of classification labels on perception and judgment. Whereas labels may summarize differences between grouped objects along a continuum, it has the potential to both exaggerate differences between the boundaries, and to obscure differences subsumed under the category labels. Part of the appeal of the peasant joke is the extreme role given to category labels, which is treated as determinative.

Category labels have been investigated extensively and labeling has been shown to affect cognitive processes in many different ways. We will first review relevant research on labeling effects on perception; then we will provide a brief overview of research on the effects of naming on categorization and memory. Finally, we will turn to the present research.

BACKGROUND

In an early study on the role of verbal labels on perception, Tajfel and Wilkes (1963) asked whether the application of category labels would distort the perception of simple objects, by making the objects within a category appear more similar to one another than they actually were (assimilation), and by making objects on the opposite side of a category boundary appear more different than they actually were (accentuation or contrast). In a series of original, elegant studies, participants were presented with a series of eight vertical lines, graded in length and differing by a fixed ratio. Subjects estimated the length of each of the eight lines under one of two main conditions: (1) a random label condition, in which the letter A or B was randomly paired with each line, or (2) a meaningful label condition, in which the letter A was paired only with the four shorter lines, while the letter B was paired only with the four longer lines.[1] Although the authors claimed support for both intercategory accentuation and intracategory assimilation, their data at best supports only one particular form of between-category accentuation.

In the area of cognitive development, there is a large literature on the effects of labeling on perceptual learning and discrimination in children (e.g., Katz, Karp, & Yalisove, 1970; Robinson, 1955; Tighe & Tighe, 1966) and on the effects of naming on categorization (Waxman & Gelman, 2009) showing that, from a very young age, words draw our attention to object categories (cf. Lupyan, 2008a). Even for infants, the use of nouns, or noun forms, to name objects appears sufficient to develop implicit categories (Waxman & Gelman, 2009; Waxman & Markow, 1995). Gelman and Heyman (1999) demonstrated the greater power of nouns over verb predicates in making behavioral inferences. They found that statements such as "Jane is a carrot eater" led to stronger behavioral inferences than "Jane eats carrots whenever she can," even though the latter conveys a more extreme behavioral propensity. Similarly, Carnaghi and colleagues (Carnaghi et al., 2008) investigated the inductive potential of nouns versus adjectives, showing that nouns have a more powerful

1. In the original paper, the authors also included a no label condition, but this was combined with the random label condition as they both showed the same pattern of results.

impact on person perception than adjectives (e.g., nouns imply more essentialism of congruent preferences). Nouns, adjectives, and verb predicates differ in a number of ways. Nouns may be more likely to encourage thinking based on discrete categories, with prototypes that are likely to be extreme examples—"paragons" in Lakoff's terms (1987), rather than an "average" of category members. Carnaghi and colleagues suggest that names, more than adjectives, inhibit alternative classifications and imply essentialistic attributions. More generally, this may be an interesting case where language usage plays an important role in defining the "reality" of category labels, even where the regions defined by the category label are somewhat arbitrary. Walton and Banaji (2004), using a modified form of Gelman and Heyman's methods found similar results with college students, even when participants themselves generated the noun or predicate phrases.

In the cognitive literature, there have been a number of lines of research showing that conceptual categories often can influence perceptual judgments, as a supplement to the more traditional finding that perceptual features can activate higher cognitive processes. Goldstone (1995), in the domain of color perception, has shown that color matching judgments during simultaneous presentation of stimulus and target are influenced by the hue of the conceptual category to which the stimulus object belongs. Lupyan demonstrated the importance of conceptual categories in a visual search paradigm, and in separate studies showed that recognition memory for visual objects was influenced by presented objects' goodness of fit to their conceptual category (2008a). For example, drawings of chairs typical of the category "chairs" were more likely to be falsely described as "new" rather than "old" in a recognition memory paradigm than were atypical members of the category. Typical chairs activated the generic category which may then have generated features (e.g., armrests) that were not present in the initial drawing, whereas presented atypical drawings, by virtue of not activating the superordinate category, were more likely to be recognized as "old." Even though the perceptual features associated with a typical category member can activate the superordinate category label, that label could then impair memory for the examplar that did not possess all the features of the activated category labels make objects appear more typical and, as a consequence, are judged as more typical in the presence of labels (cf. Lupyan, 2008a).

Goldstone, Lippa, and Shiffrin (2001) predicted contraction and expansion effects (what we refer to as assimilation and contrast) following a category learning experiment involving male faces. Using two different types of similarity judgments, they distinguished between similarity judgments that could be a result of experimenter demand, and those that reflected true differences in cognitive representation. Direct similarity judgments between two faces—which could be attributed in part to experimenter demand—tended to show assimilation effects, whereas similarity judgments mediated by a noncategorized face, and presumably representing nonstrategic judgments, showed primarily contrast effects.

THE PRESENT RESEARCH

The goal of the present research is not simply to replicate previously demonstrated assimilation and contrast effects of category labels, although we note it has been surprisingly difficult to demonstrate both these effects in the same experiment

(e.g., Goldstone, 1994; Goldstone, et al., 2001; McGarty & Turner, 1992; Richardson, 1991; Tajfel & Wilkes, 1963; but see Harnad, 1987; Livingston, Andrews, & Harnad, 1998; Rothbart, Davis-Stitt, & Hill, 1997). Our goal is to examine the effects of a label's semantic content on the strength of categorization effects, and to do so using a paradigm in which participants are simultaneously given information both about an object's position along a meaningful continuum, as well as the classifying labels associated with regions of that continuum. As both continuous and categorical information is presented simultaneously, the content of the classifying labels can serve to emphasize either the discreteness or the continuity of the information presented. As is clear from the developmental literature cited earlier, almost any noun-like label may promote categorical thinking and may promote the focus on communalities among objects sharing the same labels. In the research presented here our focus is on the full classification system, involving multiple labels, and the degree to which the labels accentuate or diminish categorical judgments. In our view, there are two general factors that are central to understanding the effects of such category labels. The first is the general social context in which the labeling occurs, and the second is the semantic content of the classifying labels. These two issues will be discussed in turn.

LABELING SOURCE

When an outside source applies a label to a collection of objects, it is assumed that the source has expert knowledge, and therefore that label has meaning above and beyond that which is conveyed by the underlying continuum. Putnam (1975, 1988) noted that labels are applied to objects precisely because they impart expert knowledge and convey useful information. This is most apparent during language learning, when adults correct a child's categorization errors by supplying an alternative label (e.g., "penguins are birds, not fish"). In psychology experiments, the experimenter often supplies a label for a collection of objects, and it is reasonable for the participant to assume that the experimenter has special knowledge that is unknown to the participant, a point that is implicit in the Strategic Model put forth by Goldstone et al. (2001). We address this question in Experiment 1 by creating a condition in which the labels are generated by the participant him- or herself, and compare it to a condition in which the labels are generated instead by an external source (i.e., peer students).

LABEL STRENGTH

Apart from the social context in which labeling occurs, the semantic content of the classifying system should also affect the strength of categorization effects. It is our view that labels vary in degree of strength, or in their ability to transform an underlying continuum into a set of apparently discrete categories. Allport, in *The Nature of Prejudice* (1954), alluded to category strength when discussing the linguistic properties of labels. He referred to "labels of primary potency" as "exceedingly salient and powerful . . . tend[ing] to prevent alternative classification, or even cross-classification. . . . These symbols act like shrieking sirens, deafening us to all finer discriminations that we might otherwise perceive" (Allport, 1954, p. 179).

Allport was referring to highly affectively toned categories, such as "communist" in his day, "socialist" in our time, or the racial epithets of both periods. It is worth noting that his phrase, "deafening us to all finer discriminations," may be one way of describing an especially strong form of within-category assimilation.

We consider the strength of a classification system to be based on three interrelated characteristics: differences in affective strength across categories, permeability of boundaries between categories, and judged discreteness of the continuum. All of these dimensions can be derived from Campbell's (1958) analysis of Gestalt principles relevant to the perception of a social group as a thing or entity. The principles of similarity and good continuation are of particular importance, where similarity is defined as affective similarity, and good continuation is defined in two ways: solidity of group boundaries (i.e, difficulty of moving from one category to another), and the divisability of the underlying continuum into discrete categories.

The *first*, affective differences, simply recognizes that affective content is a paramount feature of social categories, and differences between categories in affective valence or extremity constitute one central way of creating implicit boundaries between categories. Although Allport's "labels of primary potency" refer to labels of extreme affective polarity, less affectively polarized categories can also function to emphasize category differences. For stimuli involving line length in Tajfel and Wilkes's (1963) research it is hard to imagine category labels that differ appreciably in degree of affect, but other continua could involve meaningful affective differences for adjacent categories. It would be possible to use labels that convey small differences in affect between adjacent categories (e.g., "not attractive" vs. "attractive") or large differences (e.g., "ugly" vs. "beautiful").

The *second* aspect of label strength concerns the permeability/impermeability of boundaries, or the ease with which a stimulus object can in principle move from one category to another (cf. Campbell, 1958). The ease with which a stimulus could change from being, say, unpleasant to pleasant is probably greater than that of moving from ugly to beautiful. Some category boundaries are considered virtually impermeable (e.g., between women and men), while others are quite permeable (e.g., "average" to "above average" blood pressure).

The *third* index of judged discreteness concerns the degree to which a set of labels encourages continuous versus categorical thinking. In the examples above, the use of comparative labels (shorter/longer) is more likely to encourage continuous thinking, while noun phrases such as "carrot-eater" are more likely to encourage categorical thinking. Again using the Tajfel and Wilkes (1963) research on line length as an example, would there be a difference between a condition in which the labels were "short" versus "long" (what the labels "A" and "B" actually implied), and one in which they were "shorter" versus "longer." In the former case, the labels imply discreteness and in the second they imply ordinality, which should reinforce the idea of continuity. Thus, within-category assimilation and between-category accentuation should be stronger with the former than the latter pairs of labels.

We stated earlier that the three aspects of category strength were interrelated and to be viewed as a composite, rather than as independently manipulable. For example, as differences in affect across categories become greater, the apparent permeability between categories decreases, and perceived "discreteness" increases. The goal of this set of experiments is not to separate out the independent con-

tributions of these interrelated constructs, but to see if we can find evidence of differential categorization processes through the use of "weak" versus "strong" category labels.

VISUAL JUDGMENT PARADIGM

Participants were presented with frontal silhouette drawings of women, ordered along a continuum of ponderosity (ratio of weight to height) from very thin to very heavy in a two-phase experiment (see Figure 1 for the complete set of silhouettes). In the first phase, all participants were presented with the silhouettes placed along the ponderosity continuum with no category boundaries or labels and asked to judge the similarity of pairs of silhouettes, as well as estimate weight (in pounds) for each silhouette. In the second phase, the presence of category boundaries and labels was introduced and, depending on the experiment, the source, strength, and/or consequences of the labels were experimentally varied, after which the same judgments made in Phase 1 were repeated.

Apart from the goal of assessing the strength of category labels, this paradigm addressed some limitations of previous research. First, previous research relied exclusively on either global judgments of similarity (e.g., Rothbart et al., 1997) or on absolute judgments (Krueger & Clement, 1994; Tajfel & Wilkes, 1963). There is reason to expect that unanchored, global judgments may behave differently from anchored, absolute judgments (cf. Kobrynowicz & Biernat, 1997), and in this research both relative and absolute measures are included and assessed in the same paradigm and with the same stimulus material. Second, the stimuli used were familiar to participants, present at the time of judgment, and the judgments are not based on memory.[2] Third, judgments of body type and absolute weight were familiar, and at least in the latter case, relatively unambiguous.

The goal of the present research, then, is to examine the effects of category labels on the judgments of affectively laden, visual stimuli within and across category boundaries, using a particularly stringent experimental paradigm in which the stimuli are familiar, relatively unambiguous, present at the time of judgment, and the judgment task itself is familiar and objectively based. The research addresses two substantive issues. First, will there be labeling effects on categorization even when the source of the label is the participant rather than an external source? And second, does the "strength" of the classifying labels influence the magnitude of assimilation and accentuation?

OVERVIEW OF EXPERIMENTS

Experiment 1 manipulated the presence/absence of category boundaries and category labels, as well as the source of labels (self-generated vs. peer-generated) to separate the effects of category boundaries from those of category labels, as well

2. In line with this argument, Corneille and colleagues (Corneille, Klein, Lambert, & Judd, 2002) replicated Tajfel and Wilkes's (1963) accentuation effects only when using judgment scales that were unfamiliar to the participants (i.e., inches for European participants and centimeters for U.S. participants) and therefore, relatively ambiguous.

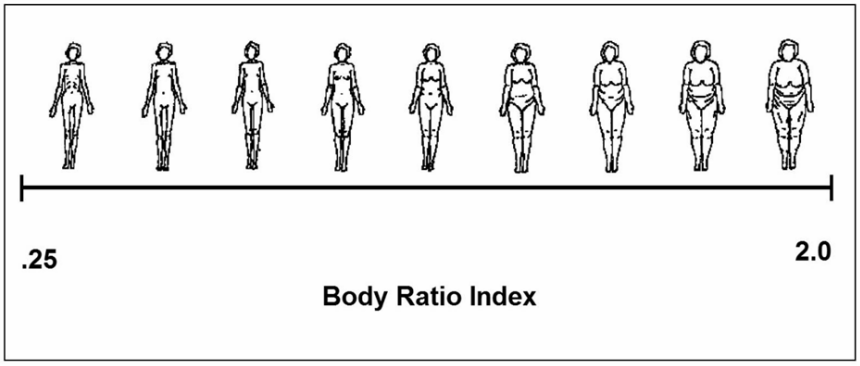

FIGURE 1. Set of silhouettes images used as stimuli. Adapted from Furnham and Alibhai (1983).

as to test the effect of different sources of labels (self vs. peer). Since there was an effect of both self- and other-generated labels, the next experiments attempted to assess possible explanations for these results. Experiment 2 compared the nature of the source of labels (peer vs. expert) independently of the strength of the labels (weak vs. strong) to test whether expertise and strength affect our judgments. Finally, Experiment 3, examined another aspect of social context, the "pragmatics" of labeling, and assessed whether labels applied for a practical (medical) purpose had different effects than labels applied for descriptive purposes only. In this last experiment, the ostensive source of the labels was medical experts, whose labels varied in strength and pragmatic implications.

EXPERIMENT 1

OVERVIEW

This experiment examined whether the presence of category boundaries and verbal labels, and the source of the verbal labels, affected the perception of category members. The participants were randomly assigned to one of four conditions that differed from one another only in Phase 2 of the Visual judgment paradigm: (a) *No category/no labels condition:* the continuum and task presented in Phase 2 is identical to that of Phase 1; (b) *Category/no labels condition*: the continuum during Phase 2 is divided into three categories of equal size, but with no verbal labels; (c) *Category/self-generated label condition:* the continuum is again divided into three categories and described with three labels generated by the participant her- or himself; and (d) *Category/peer-generated labels condition:* the continuum is divided into three categories and described with 3 labels ("anorexic," "normal," and "obese"),[3] with participants led to believe that the labels were generated by their peers.

According to accentuation theory (see Eiser & Strobe, 1972; Tajfel, 1969), perceived differences between stimuli in different categories should be accentuated (contrast effect) and differences between stimuli in the same category should be minimized (assimilation effect). Following these predictions: (1) people should show contrast effects when category boundaries and verbal labels (both self- and peer-generated) are present compared to the simple continuum or to the simple

continuum with category boundaries present; (2) when labels are present (either self- or peer-generated), the perceived "strength" of the labels (as judged by the participants themselves) should correlate with the magnitude of assimilation and contrast effects.

In ordinary discourse, when labels are associated with a class of objects, it is assumed, often implicitly, that labels reflect expert or general knowledge and therefore the labels add substantive meaning to the classified objects (cf. Putnam 1975, 1988). For this reason, the condition involving self-generated labels is of particular importance because the source of the label is the self—rather than an outside source—and thus the information available to the participants in Phase 1 and 2 are the same. The use of peers as an external source of labels was chosen since peers, by definition, have the same general level of knowledge as the participants, but nonetheless constitute a source external to the self.

METHOD

Participants

Two hundred thirty-three students (141 females) participated in this computer-based experiment as partial fulfillment of a research requirement for an Introductory Psychology course.

Stimulus Material

The stimulus material was a set of 9 female silhouettes, frontal view, at roughly equally spaced intervals along a continuum ranging from very thin to very heavy (see Figure 1; adapted from Furnham & Alibhai, 1983). The continuum and the silhouettes were selected to allow the continuum to be divided into 3 equally spaced regions with 3 silhouettes in each region. This criterion was used to achieve categories of equal size and equal number of exemplars.

Procedure

Participants were seated in front of a personal computer where they could self-administer the experiment. Standardized instructions presented on the screen described the experiment as one involving social perception. The participants were asked to look at "silhouettes of the female body, varying in the ratio of weight to height, with some silhouettes thinner and some others heavier in shape." The silhouettes were described as:

> "Varying along a single continuum called the Body Ratio Index (or BRI), which is based on a complex formula comparing body weight to height. All of the silhouettes to be presented fall between the BRI of .25 (the thinnest extreme) and 2.00 (the heaviest extreme)."

3. This set of labels was rated as strong based on the three dimensions described previously.

The experiment consisted of two phases. In Phase 1, participants viewed 15 pairs of silhouettes in random order.[4] The two silhouettes of each pair were presented together at their corresponding positions on the continuum. That is, when judging the similarity between silhouette 1 and 2, for example, participants were presented with the whole continuum and the two silhouettes standing in the first and second position from the left (see Figure 1). Participants judged the similarity between the two silhouettes on a 9-point Likert scale (from "not at all similar" to "extremely similar") on three different dimensions ("personality," "life style," and "body type"), with the higher score indicating higher similarity. After participants enter the rating for the target pair, the silhouettes disappear and the next trial started with the subsequent pair presented on the continuum. The three different dimensions of similarity (body type, life style, and personality) were chosen to allow a set of different ratings on the same stimulus targets (same pair). These three dimensions were included as a way of defining specific aspects of similarity, rather than using a single and general measure of similarity. After the similarity judgments, participants were asked to estimate the weight of each of the nine silhouettes, presented individually in random order. Each silhouette was presented together with the continuum in its correspondent position (e.g., silhouette 1 was presented on the first position on the left). Between Phase 1 and Phase 2, participants completed an unrelated task, which took approximately 10 minutes. In the second phase, participants completed the same measures as in Phase 1, but the presence of category markers and category labels was systematically varied between subjects, as follows:

No Category/No Labels Condition (i.e., Control). The same continuum as in Phase 1 was presented again, with no category boundaries or labels. The purpose of this condition was to assess the effect of repeating the same judgments, providing a baseline for comparison.

Category/No Labels Condition. Participants in Phase 2 were presented with a continuum divided by tick marks into three equally spaced regions, but with no labels attached to the regions. This condition provided the necessary control to test the effects of categories without any verbal labels.

Category/Self-Generated Labels Condition. Participants in Phase 2 were presented with the continuum divided into three regions (as in the category boundaries condition) and they were asked to generate their own labels for each region. Participants in this condition were presented with the continuum representing Body Ratio Index (BRI) and with the silhouettes from each of the three sections of the BRI. In addition, they were asked to type a "one or two word label or description that you [the participant] feel describes these body types." The three sections were presented in random order. The entries were recorded and each participant performed the remaining part of the experiment viewing his/her own labels attached to the continuum.

4. Out of the 36 possible combinations of the nine silhouettes, only the 15 pairs were considered in which the distance between the two was 1 or 2 units (e.g., silhouette number 1 and 2, and number 1 and 3, respectively in Figure 1). This selection optimized the number of pairs to be presented and the information collected and allows to have both distances involving within- and across-boundary comparisons.

Category/Peer-Generated Labels Condition. The continuum divided into the same three regions was presented with the attached labels: "anorexic," "normal," and "obese." The participants in this last condition were told that:

> A random sample of undergraduates . . . were presented with these silhouettes, and they were asked to come up with labels or descriptions. . . . The labels most commonly used to describe the three sections are, from left to right, "anorexic," "normal," and "obese."

For each of these four conditions the same similarity judgments and weight estimations were obtained in Phase 1 and Phase 2. After completing Phase 2, participants who were presented with labels (self- and peer-generated) were also asked to rate the labels he/she saw on three dimensions: (a) "Valence": degree of Positivity/Negativity of each category label (7-point likert scale from –3 "Extremely Negative" to +3 "Extremely Positive"; (b) "Categoricalness": degree to which the 3-label set is categorical instead of continuous; and (c) "Movement": how easy it would be to move from one category to another. These three questions allowed an assessment of the three components related to "strength" of category labels (cf. Allport, 1954). At the end of the computer task, the participants were thanked and debriefed.

RESULTS

Overview of Design and Data Analysis

In Experiment 1, all participants provided both similarity judgments between pairs of silhouettes and weight estimates for individual silhouettes, on two occasions. In Phase 1, judgments were made for the nine silhouettes presented along an uncategorized, unlabeled continuum. In Phase 2, participants provided the same judgments but after experiencing one of four possible conditions: (1) an uncategorized, unlabeled control condition (identical to Phase 1), (2) a categorized, unlabeled control condition, (3) a categorized experimental condition, in which participants generated their own labels for the three categories, and (4) a categorized experimental condition in which labels for the three categories were putatively generated by peers. The basic prediction was that judgments of silhouettes should show greater similarity within category, and greater differences between categories, for the labeled experimental conditions than for the control conditions. For each of the two dependent measures, Phase 1 and Phase 2 judgments were treated as within-subjects, repeated measures, and the predicted effect would be evinced by an interaction between experimental conditions and time (Phase 1 vs. Phase 2). In addition to the basic prediction, there was also the prediction that the "other-generated" labels would have stronger effects than would the "self-generated" labels. No strong prediction was made regarding differences between the two control conditions.

Similarity Judgments

Data Reduction. Each participant made three similarity judgments for each of 15 pairs, in both Phase 1 and in Phase 2. The three different dimensions of similar-

ity (physical, personality, and lifestyle similarity) were highly correlated, with an average within-subject correlation among the three dimensions of $r = .60$ for Phase 1; and $r = .71$ for Phase 2. Because of the high correlations, the three similarity judgments were averaged to create a single similarity measure for each pair of silhouettes. The 15 pairs varied according to the distance between the members within a pair (1 vs. 2 units), and whether the members of the pair existed within the same or different categories, resulting in four different conditions: 1 unit apart/across-category boundary, 2 units apart/across-category boundary, 1 unit apart/within-category boundary, and 2 units apart/within-category boundary. Similarity judgments within each of these four subsets of pairs were averaged, and then the 1 unit and 2 unit distance conditions were further averaged to yield a single within-category similarity score and a single across-category similarity score.[5] The two scores were computed for the judgments made during both Phase 1 and Phase 2, and used as repeated measures in a within-subjects design.

Data Analyses. In a separate analysis of variance, no significant difference was found between the two no label control conditions (No category/no labels and Category/no label condition), and these two conditions were combined into a single combined control/no labels condition.[6]

A mixed 2 x 2 x 3 x 2 ANOVA was conducted, where the first 2 factors were within subjects (*Boundary*: within vs. across; *Time*: "Phase 1" vs. "Phase 2") and the last two were between-subject factors (*Labels*: "Combined control/no labels" vs. "Category/self-generated label" vs. "Category/peer-generated label"; *Gender* of the participants). The means and standard errors are presented in Figure 2. For ease of comprehension all the figures represent the factor time as a different score.[7]

The ANOVA revealed a strong interaction between *Boundary* and *Time*, $F(1, 227) = 63.82$, $p < .001$, $\eta^2 = .22$, indicating the perceived similarity was greater for within- than between-category stimulus pairs in Phase 2 than in Phase 1. Most importantly, however, the greater differences for within- than between-category comparisons in Phase 2 was itself related to *Labels*, as indicated by the expected three-way interaction between *Labels*, *Time*, and *Boundary*, $F(2, 227) = 9.97$, $p < .001$, $\eta^2 = .08$. *Gender* of the participants show only a main effect on judgment, $F(1, 227) = 4.70$, $p < .05$, $\eta^2 = .02$, indicating that female participants rated the pairs, in general, as more similar than male participants did, so this factor was not considered any further in the analyses.

To separate the effects of within-category similarity (assimilation) from across-category similarity (contrast effect), 2 x 3 ANOVAs (*Time* x *Labels*) were run separately on within- and across-boundary pairs.

For the within-category pairs, the expected 2-way *Time* x *Labels* interaction was highly significant, $F(2, 230) = 5.52$, $p = .005$, $\eta^2 = .05$. Planned contrasts showed the two label conditions as having greater similarity over time than the combined

5. Analyses including the factor distance (1 unit vs. 2 units apart) did produce parallel results with the obvious addition of the consistent main effect of distance: pairs that are 1 unit apart are perceived consistently more similar than pairs 2 units apart.

6. Across three experiments, there were never any significant differences between the two control conditions based on similarity judgments or absolute weight estimates, and thus in the remaining of the study the two conditions will be always treated as one combined control condition.

7. In this case, indexes of Phase 1 were arbitrarily subtracted from those of Phase 2 so that a positive value would represent an increase of similarity between the two silhouettes from Phase 1 to Phase 2. It should be noted that ANOVAs based on difference scores showed virtually identical statistical decisions to ANOVAs using time as a repeated measure.

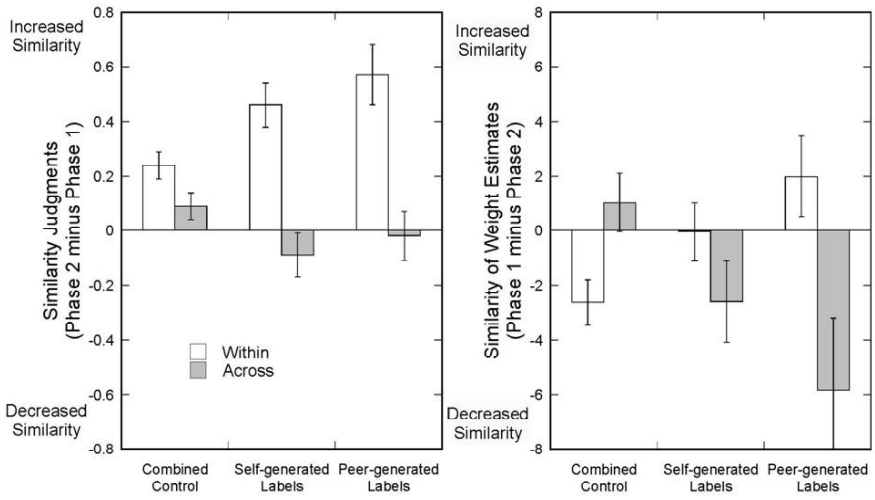

FIGURE 2. Experiment 1: Difference in similarity judgments (left panel) and weight estimates (right panel) from Phase 1 to Phase 2 for within- and across-boundary pairs by Labels condition. Error bars represent standard error of the mean. Positive scores represent increased similarity (i.e., assimilation effects) while negative scores represent decrease similarity (i.e., accentuation effect). Putatively peer-generated labels were: anorexic, normal, obese.

control, $F(1, 230) = 11.01$, $p = .001$, $\eta^2 = .05$, but the label conditions did not differ from each other ($F < 1$).

For across-category pairs, the expected interaction between *Time* and *Labels* did not reach the standard level of significance, $F(2,230) = 2.17$, $p = .12$, $\eta^2 = .02$. Planned contrasts showed that labels reduced the similarity of across-category pairs (accentuation) from Phase 1 to Phase 2 compared to the control, $F(1,230) = 3.26$, $p = .07$, $\eta^2 = .014$, but again did not differ from each other, ($F < 1$). These results provide clear support for intracategory assimilation and only marginal support for intercategory accentuation effects.

Absolute Estimation of Weight

Data Reduction. Absolute weight estimates allowed the examination of assimilation and contrast effects in a way that is exactly parallel to the analysis of similarity judgments. Using as a reference the same 15 pairs of silhouettes rated for similarity, the difference in pounds between the estimates of the two silhouettes of each pair was computed. Arbitrarily, the weight estimate of the lighter silhouette in the pair was subtracted from the estimate of the heavier one. In this way, it was possible to create a difference score for each pair that would be comparable to the similarity judgments. Difference scores were then aggregated according to the different subsets of pairs used for similarity judgments (i.e., 1 unit apart/across-category, 2 units apart/across-category, 1 unit apart/within-category, and 2 units apart/within-category). Differences in estimates within each of these four subsets of pairs was averaged, and subsequently, the two across-category boundaries scores and the two within-category boundaries scores were averaged. In this way, each participant was represented again by four distinct scores: a single within-category "similarity" score and a single across-category "similarity" score calculated both for Phase 1 and Phase 2. The four scores (across and within similarity in Phase 1

and 2) were used as a repeated measures, within subjects factor as was done for similarity judgments. Due to the nature of the absolute estimates, the smaller the differences between a pair of silhouettes, the greater the similarity.

Data Analyses. As for the similarity judgments, a 2 x 2 x 3 x 2 ANOVA was conducted on the absolute weight estimates, with two within-subject factors (*Boundary*: Within vs. Across; *Time*: "Phase 1" vs. "Phase 2") and two between-subjects factors (*Labels*: "Combined control/no labels" vs. "Category/self-generated label" versus "Category/peer-generated label"; *Gender* of the participants). The means and standard errors for the three conditions are presented in Figure 2.

There was a main effect of *Boundary*, $F(1, 227) = 55.48$, $p < .001$, $\eta^2 = .20$, showing that in general across-boundary pairs are estimated to be closer in weight ($M = 24.53$) than the within-boundary pairs ($M = 30.32$).[8] There was also an effect of *Time*, $F(1, 227) = 13.21$, $p < .001$, $\eta^2 = .05$, indicating that, in general, silhouettes are estimated to weight more during Phase 2 ($M = 28.74$) compare to Phase 1 ($M = 27.28$). The two-way interaction between *Boundary* and *Time* showed the same pattern of increased similarity (decreased differences) for within- than across-category pairs from Phase 1 to Phase 2, $F(1, 227) = 4.64$, $p = .032$. More importantly, the predicted three-way interaction between *Labels*, *Time*, and *Boundary* was significant, $F(2, 227) = 7.51$, $p = .001$, $\eta^2 = .06$. Parallel to the findings for direct similarity judgments, *Gender* of the participants was significant only as a main effect, $F(1, 227) = 6.97$, $p < .01$, $\eta^2 = .03$, showing that in general female participants perceived the silhouettes to be closer in weight than male participants did so this factor was again not considered any further in the analyses.

Again to separate the effects of within-category assimilation from between-category accentuation, 2 x 3 (*Time* X *Labels*) ANOVAs were run separately for within- and across-boundary judgments. For within-boundary pairs, the expected two-way interaction *Time* x *Labels* was significant, $F(2, 230) = 4.41$, $p = .013$, $\eta^2 = .04$. Planned contrasts showed that within-category pairs were estimated as closer in weight (assimilation), from Phase 1 to Phase 2, in the label conditions than in the combined control condition, $F(1, 230) = 8.59$, $p = .004$, $\eta^2 = .04$, and as with similarity judgments, the label conditions did not differ from each other, $F(1, 230) = 1.27$, *ns*.

For across-category pairs, the expected interaction between *Time* and *Labels* was also significant, $F(2,230) = 4.66$, $p = .01$, $\eta^2 = .04$, with differences between silhouettes over time varying by condition. Planned orthogonal contrasts showed that the two label conditions significantly increased the distance in estimation of across-category pairs (accentuation) from phase 1 to phase 2 compared to the control conditions, $F(1,230) = 8.98$, $p = .003$, $\eta^2 = .04$ but again did not differ from each other, $F(1,230) = 1.58$, *ns*. These results provide clear support for both intracategory assimilation and for intercategory accentuation

8. Although the main effect indicating greater similarity for across- rather than within-pairs appears counter to the hypothesis, it is not. It is due to the different nature of the within- and across-pairs. First, within-boundary pairs (1-2, 2-3, 1-3, 4-5, 5-6, 4-6, 7-8, 8-9, 7-9) are different from the across-boundary pairs (2-4, 3-4, 3-5, 5-7, 6-7, 6-8), and the former includes the extremes of the scale (silhouettes 1 and 9), while the across-pairs do not, leading to stronger perceived differences for the within- than across-pairs. Most importantly, the presence of these baseline differences are assessed in the Time 1 control condition and the key theoretical issue is how the measures of similarity change as a result of the superimposition of category labels from Time 1 to Time 2. Given the nature of the different within- and across-pairs, the main effect of boundary (present in all three experiments) is of no theoretical relevance.

Effect of the Strength of Labels

To assess the perceived strength of category labels, the three scales were aggregated into a single *strength index*. The strength of a set of three labels was considered to increase with: (1) the average evaluative difference between each one of the two external categories and the central category; (2) the judged "categoricalness" of the set of labels; and (3) the (averaged) difficulty in moving from one category to another. For "Valence," participants indicated how positively or negatively he/she regarded each label, on a 7-point likert scale ranging from –3 ("Extremely Negative") to +3 ("Extremely Positive"). For "Categoricalness" participants indicated to what degree the three labels reflected differences along a continuum or between discrete categories on a 7-point likert scale ranging from "Extremely Continuous" to "Extremely Discrete." For "Movement," participants judged, for each pair of adjacent categories, how easy or difficult it was to switch from one body type to another, using a 7-point Likert scale ranging from 1 ("Extremely easy") to 7 ("Extremely difficult"). This averaged index was then correlated with the different dependent measures (e.g., similarity judgments for within- and across-category pairs), predicting that the greater the perceived strength of the labels (i.e., strength index), the greater the assimilation and contrast effects on both similarity and weight estimates.

Forty-two participants from the "Category/self-generated label" condition and 22 from "Category/peer-generated label" condition provided data for the correlations. The strength index correlated significantly with each one of the 2 dependent measures. First, the strength index correlated with the difference score (Phase 1 minus Phase 2) based on estimates of weight for the across-boundary pairs (r = -.31, p = .012), indicating that the stronger the labels, the larger the boundary accentuation effects. In addition, the strength index correlated positively with similarity judgments for within-category pairs (r = .32, p = .01), with stronger labels associated with greater similarity within each category.

It should also be noted that the two label conditions differed in perceived strength, with peer-generated labels (M = 4.2) rated as stronger than self-generated labels (M = 3.6), $t(62)$= 2.78, p = .007. However, even when the two conditions were considered separately the same pattern of correlations was found with slightly reduced significant levels, indicating that these results are not artifact of systematic differences between the conditions.

DISCUSSION

Compared to the combined control condition, both verbal label conditions affected judgments and did so for both the more abstract and relative judgment of similarity as well as for the familiar and absolute judgment of weight. For similarity judgments between pairs of silhouettes, there was strong evidence of within-category assimilation, but no clear evidence of accentuation at the boundaries between categories. For judgments of absolute weight of individual silhouettes, there was evidence of both within-category assimilation and intercategory accentuation.

The weak evidence of intercategory accentuation using the similarity measure will be discussed more extensively in the general discussion. Strong evidence for assimilation was present for both measures and evidence for contrast was present for the absolute judgment of weight. In comparison to previous research (i.e., Rothbart et al., 1997), similarity judgments did not show any effect of the presence of category boundaries without labels. This is probably due to features of the paradigm. In fact, given the familiarity of the visual stimuli, the repeated measure design in which each participant provides the same judgments in the absence and presence of verbal labels, the presence of the stimuli at the time of judgment, and the familiarity of the measures, the evidence for labeling effects on categorization processes appears particularly strong. Perhaps the most important finding, however, concerns the self-generated labels. In this condition participants generated their own category labels, and nonetheless were subsequently influenced by their own labels in judging the silhouettes. This finding weakens the argument that outside or expert knowledge is a necessary condition for labeling effects, at least in this experiment. It is hard to argue that self-generated labels add information to the stimulus array or change the participants' knowledge of the continuum from Phase 1 to Phase 2 in the present paradigm. Based on the correlational data, the labels judged as stronger (i.e., more evaluatively different, more categorical, and with less permeable boundaries), increased perceived within-category similarity (based on the similarity measure), and decreased similarity between silhouettes under different labels (based on weight estimates). These effects were present within both the self-generated labels condition and the peer-generated labels condition. Even if interesting and clear the correlational data only suggest that the strength of the labels may modulate the magnitude of the labeling effects. In order to experimentally test such hypothesis, strength of the labels should be manipulated while keeping constant other relevant dimension (e.g., source of the labels). Even though the strength of category differed across the two label conditions, Experiment 2 was designed to test the strength hypothesis more directly by experimentally manipulating the strength of the labels.

EXPERIMENT 2

OVERVIEW

In Experiment 1 we showed that the simple presence of labels affects the perception of objects (self-generated condition) and that source of the labels (self vs. peer) did not appear to matter: two nonexpert sources (self and peer) do not show significantly different effects. Although one could expect stronger categorization effects by peer than by self, those effects were not found. However, it is nonetheless possible that labels generated by an expert source might produce stronger effects than labels from a nonexpert source. Thus, label strength was explicitly varied in this experiment. The goal of Experiment 2, then, was to explicitly assess the effects of expertness of source and strength of labels by direct experimental manipulation.

METHOD

Participants

Two hundred-two students (151 females) participated in this computer-based experiment as partial fulfillment of a research requirement for an Introductory Psychology course.

Stimulus Material, Design, and Procedure

The stimulus materials and procedure used were the same as in Experiment 1, unless otherwise specified. The strength of the labels was manipulated, based on a pretest, using two different sets of labels: weak labels ("below average," "average," and "above average"), and strong labels ("anorexic," "normal," and "obese"). In addition to the strength of the labels, the expertise of the source of the labels was manipulated (peer vs. expert). The peer-generated labels were introduced as in Experiment 1. In the expert-generated labels condition, participants were told that:

> A random sample of Doctors from the American Nutritionists Association (ANA) was presented with these silhouettes, and they were asked to come up with labels or descriptions to describe each of these three sections of the continuum. The labels most commonly used to describe the three sections are, from left to right [. . .]

In Phase 2, the nature of the continuum on which the stimuli were presented was systematically varied between subjects. Each participant was randomly assigned to one of the following experimental conditions:

Combined No Label Control Condition. The participants during Phase 2 were presented with the continuum as in Phase 1 (or with a continuum divided into three regions), without category labels.

Peer-Generated Weak Category Label Condition. The source was a peer, and the continuum in Phase 2 was divided into three regions with three "weak" labels attached: "below-average," "average," and "above-average."

Peer-Generated Strong Category Label Condition. The source was a peer, and the continuum was the same as the previous condition, but with strong labels: "anorexic," "normal," and "obese."

Expert-Generated Weak Category Label Condition. The source was a physician specializing in nutrition, and the continuum in Phase 2 was divided into three regions with three "weak" labels attached: "below-average," "average," and "above-average."

Expert-Generated Strong Category Label Condition. The source was a physician specializing in nutrition, and the continuum was the same as in the previous condition, but with strong labels: "anorexic," "normal," and "obese."

The basic design of the experiment was a 2 x 2 factorial design (*Strength*: weak vs. strong x *Expertise*: peer vs. expert), with two nonorthogonal control conditions: (1) a no category/no label condition, and (2) a category/no label condition. Due to the complexity of the 2 x 2 factorial design with the addition of two nonorthogonal

control conditions, for the remainder of the article a two-step analysis strategy is followed. First, the four label conditions are compared to the no-label combined control conditions; and second, comparisons within the four label conditions by means of a standard 2-way ANOVA are performed. This strategy will be followed for both similarity judgment and absolute weight estimates. One final difference from Experiment 1 was the absence of the intervening task.

RESULTS

Manipulation Check

At the end of the experiment, participants in the label conditions were also asked to rate the set of labels on the three aspects of category strength. A strength index was created for each participant. A *t*-test on the strength index was computed comparing the weak labels conditions against the strong labels conditions. As expected the weak labels (M = 3.61, SD = .70) were rated as less strong than the strong labels (M = 4.08, SD = .84), $t(154) = -3.77$, $p < .001$.

Similarity Judgments

For the similarity judgment, the same data reduction procedure was used as in Experiment 1. Again similarity judgments among the three different dimensions of similarity showed a high positive correlation and thus were averaged to create a single similarity measure for each pair of silhouettes (average within-subject correlations among the three dimensions were r = .60 for Phase 1, and r = .72 for Phase 2).

The design allowed the computation of a 2 x 2 x 2 x 2 ANOVA with two within-subject factors (*Boundary*: within vs. across; *Time*: "Phase 1" vs. "Phase 2") and two between-subject factors *(Labels*: combined control/no label vs. combined label; *Gender* of the participants). The means and standard errors for the five conditions are presented in Figure 3. Gender of the participants showed no effects on similarity judgments neither as a main effect nor in interaction, thus was not considered further in the analyses. First, there was evidence of an interaction between *Time* and *Boundary*, $F(1, 200) = 38.38$, $p < .001$, $\eta^2 = .16$, showing that an increase in similarity from Phase 1 to Phase 2 was greater for the within- than for the across-boundary pairs. Second and most importantly, the expected interaction between *Time*, *Boundary,* and *Labels* was also significant, $F(1, 200) = 4.00$, $p < .05$, $\eta^2 = .02$, showing overall that the labeling conditions differentially influenced within- and across-boundary judgments in comparison to the control conditions.

As categorization effects significantly differed between the label and control conditions, we now turn to the effects of expertise of the source and strength of label. As in Experiment 1, to sort out the separate effects of within-category similarity (assimilation) and across-category similarity (contrast effect), separate 2 x 2 x 2 ANOVAs (*Time*: "Phase 1" vs. "Phase 2" x S*trength*: "weak category labels" vs. "strong category label" x *Expertise*: "expert-generated" vs. "peer-generated label") were run on within- and across-boundary pairs on the label conditions only (remaining N = 157).

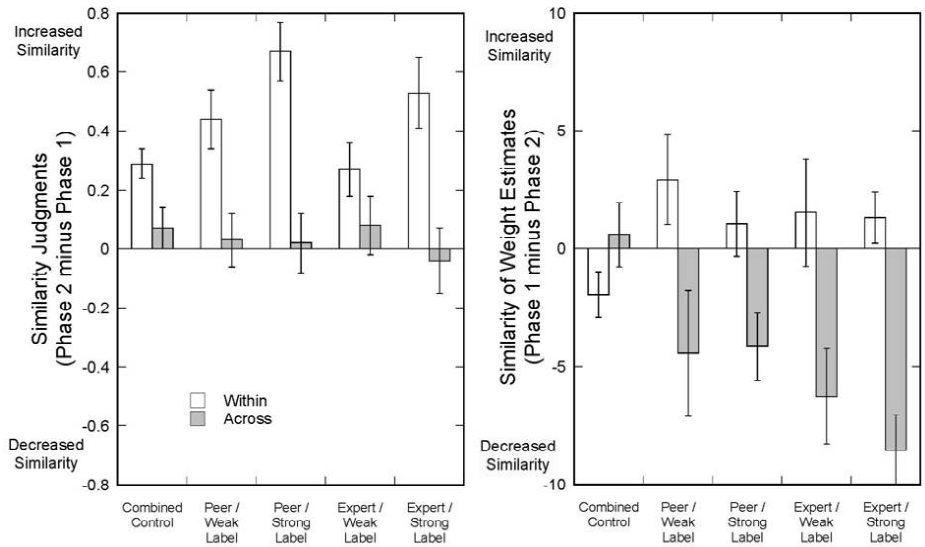

FIGURE 3. Experiment 2: Difference in similarity judgments (left panel) and weight estimates (right panel) from Phase 1 to Phase 2 for within- and across-boundary pairs by Label condition. Error bars represent standard error of the mean. Positive scores represent increased similarity (i.e., assimilation effects) while negative scores represent decrease similarity (i.e., accentuation effect). Labels were putatively generated either by peers or by expert-nutritionists. Weak labels were: below-average, average, above-average; Strong labels were: anorexic, normal, obese

For within-boundary pairs, the only significant interaction effect was between *Time* and *Strength*, $F(1, 153) = 5.51$, $p = .02$, $\eta^2 = .03$, indicating that strong labels, more than weak labels, increased within-boundary similarity from Phase 1 to Phase 2. There was no effect of expertness of source, either alone, or in interaction with time.

For across-boundary pairs, the similarity from Phase 1 to Phase 2 did not change significantly as a function of strength or expertise. These results, then, provide clear support for intracategory assimilation, but no support for intercategory accentuation.

Absolute Estimation of Weight

For the absolute estimates of weight, the same data reduction procedure was used as in Experiment 1. The design was the same as the one used for similarity judgments. The means and standard errors are presented in Figure 3. Again *Gender* of the participants showed no effects on absolute judgments neither as main effect nor in interaction, thus was not considered further in the analyses. The interaction between *Time* and *Boundary*, $F(1, 200) = 4.73$, $p = .031$, $\eta^2 = .02$, indicates that absolute weight estimates, over time, became more similar for within-category pairs than they did for between-category pairs. And most importantly, the expected interaction between *Time*, *Boundary,* and *Labels* was also significant, $F(1, 200) = 6.21$, $p = .014$, $\eta^2 = .03$, indicating the categorization effects were greater for the labeled than the nonlabeled conditions. Finally, there was a main effect of *Boundary*, $F(1, 200) = 37.78$, $p < .001$, $\eta^2 = .16$, showing that in general across-boundary pairs are

estimated to be closer in weight (M = 25.74) than the within-boundary pairs (M = 31.29).

Again to separate the effects of within-category assimilation from between-category accentuation, 2 x 2 (*Time* X *Labels*) ANOVAs were run separately for within- and across-boundary judgments. For within-boundary pairs, the expected 2-way interaction *Time* x *Labels* failed to reach standard levels of significance, $F(1, 200) = 2.67$, $p = .13$, $\eta^2 = .01$. For across-category pairs, the expected interaction between *Time* and *Labels* was significant, $F(1, 200) = 6.51$, $p = .011$, $\eta^2 = .03$. These results provide clear support for intercategory accentuation and nonsignificant support for intracategory assimilation. The second step ANOVA, assessing the effects of strength of label, and expertness of source, showed that categorization effects, evinced in the interaction between Boundary and Time, were not modulated by these factors.

Effect of the Strength of Labels

To assess the perceived strength of category labels, as in Experiment 1, the single *strength index* (aggregated across the three scales) was correlated with the different dependent measures (e.g., similarity judgments for within- and across-category pairs). Participants from the label condition provided data for the correlations (N = 157). As in Experiment 1, the strength index correlated positively with similarity judgments for within-category pairs ($r = .15$, $p = .06$), with stronger labels associated with greater similarity within each category. In this case, the correlation between the strength index and the difference score (Phase 1 minus Phase 2) based on estimates of weight for the across-boundary pairs was not significant.

DISCUSSION

This experiment shows again that labels influence the judgments of category members on both similarity judgments and absolute estimates of weight. Within-category assimilation is present for both measures (albeit not significant for absolute judgments), while across-category accentuation is present only for the absolute measure of weight (as in Experiment 1). Experiment 2 tested whether the level of expertise of the source and the strength of the labels influenced the labeling effect. The results show that when looking at the label conditions only, label strength is the only factor modulating the labeling effect for similarity judgments. Surprisingly, the effect of source expertise had little effect on judgments. Experiment 1 and 2, then, are consistent in showing that a label generated by the self, by peers, or by experts have a consistent effect on judgments, with the primary source of variation being the strength of the label and not its source. Again, as in Experiment 1, we found support for assimilation and mixed support for boundary effects (accentuation), which will be considered in the general discussion.

In summary, Experiment 2 clearly indicates that it is the strength of the label rather than its source that influences the perception of categorized stimuli. It was surprising that labels provided by medical experts had no greater influence on judgment than did labels provided by peers. One possible explanation for the lack of difference has to do with the putative implications of the categorization process.

Presumably, medical experts divide people into categories for a reason, and that reason concerns possible courses of treatment. Peers categorize for other reasons, presumably related to social acceptance or social ostracism. One possible explanation for the effects of the strength of the labels is that the implicit consequences associated with the weak and strong labels differ. To test this alternative hypothesis, the strength of the labels and the consequences of the labels were independently manipulated in Experiment 3. In this experiment the presence or absence of a category's treatment implications was explicitly varied (in a medical context), along with the strength of the labels.

EXPERIMENT 3

OVERVIEW

Experiment 3 examined the independent effects of strength of label and the consequences of being labeled. In Experiment 2, it is possible that for the expert-generated strong labels (e.g., "anorexic" or "obese"), participants implicitly assumed that a medical intervention was contingent upon labeling, whereas, social ostracism was more likely to be a consequence of peer-generated strong labels providing in both case some consequences of being labeled. For the weak labels, such consequences might have been less likely to be assumed. More generally, it is possible that one of the uncontrolled pragmatic aspects of labeling is the social implications of being labeled. In this experiment this hypothesis was tested directly by experimentally manipulating the consequences of the labels. In one set of conditions the labels were stated as having clear implications for medical treatment, while in another set of conditions the labels were described as descriptive only, with no medical implications.

METHOD

Participants

Two hundred-fifteen students (154 females) participated in this computer-based experiment as partial fulfillment of a research requirement for an Introductory Psychology course.

Stimulus Materials, Design, and Procedure

The stimulus materials and procedure were exactly as in Experiment 1, unless otherwise specified. The basic design of the experiment was a 2 x 2 factorial design (*Strength*: weak vs. strong x *Implication*: absent vs. present), with two nonorthogonal control conditions: (1) a no category/no label condition, and (2) a category/no label condition. For all the four experimental label conditions, "Doctors from the American Nutritionists Association" were described as the source of the classification and the labels. For the experimental conditions, the implications—or the absence of implications—of being a member of one of the labeled categories was explicitly stated. In the no-implications conditions, participants read that the clas-

sification and labels were "used only for descriptive purposes, they [doctors] do not base any treatments, clinical interventions or diet recommendations on a person's category membership . . ."

In the condition where the labels have implications, it was stated that doctors "use this classification both as a description and as the basis for treatments, clinical interventions, and diet recommendations . . ." For simplicity, we henceforth describe the no-implication conditions as "descriptive" in nature and the implication condition as "medical" in nature.

As in Experiments 1 and 2, all participants made the same judgments associated with Phase 1. In Phase 2, the conditions differed as follows:

Combined No Label Control Conditions. The participants during Phase 2 were presented with the continuum as in Phase 1 (or with a continuum divided into three regions), without category labels.

Weak Label No-Implication Condition. Participants were presented with the continuum divided by tick marks into three equally spaced regions, with the three "weak" labels attached ("below-average," "average," and "above-average"), with the labels portrayed as only descriptive.

Strong Label No-Implication Condition. Participants were presented with the continuum divided by tick marks into three equally spaced regions, with three "strong" labels attached ("anorexic," "normal," and "obese"), and the labels portrayed as only descriptive.

Weak Label Implication Condition. Participants were presented with the continuum divided by tick marks into three equally spaced regions, with three "weak" labels attached ("below-average," "average," and "above-average"), and the labels described as having medical implications.

Strong Label Implication Condition. Participants were presented with the continuum divided by tick marks into three equally spaced regions, with three "strong" labels attached ("anorexic," "normal," and "obese"), and the labels described as having medical implications.

As in Experiment 2 there was no intervening task.

RESULTS

Data analysis follows the same two-step strategy used in Experiment 2, where first the labeled conditions are compared with a control, and second, within the labeled conditions, the effects of a label's strength and consequences are assessed.

Manipulation Check

As in Experiment 2, participants in the label conditions were asked to rate the set of labels on the three aspects of category strength. A strength index was created for each participant. A *t*-test on the strength index was computed comparing the weak labels conditions against the strong labels conditions. As expected the weak labels

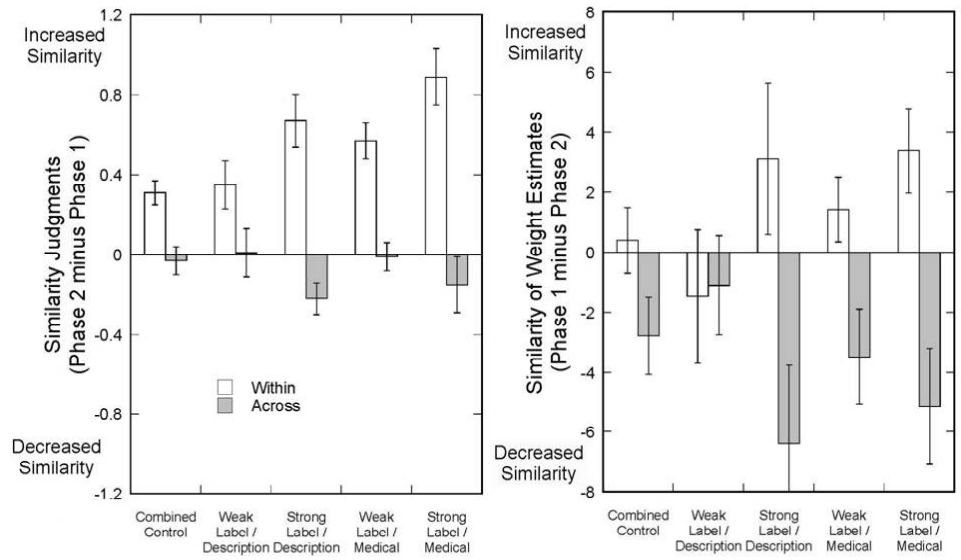

FIGURE 4. Experiment 3: Difference in similarity judgments (left panel) and weight estimates (right panel) from Phase 1 to Phase 2 for within- and across-boundary pairs by Label condition. Error bars represent standard error of the mean. Positive scores represent increased similarity (i.e., assimilation effects) while negative scores represent decrease similarity (i.e., accentuation effect). Labels were putatively generated by expert-nutritionists and had either only descriptive purposes (i.e., Description) or had clinical implication (i.e., Medical). Weak labels were: below-average, average, above-average; Strong labels were: anorexic, normal, obese

(M = 3.77, SD = .64) were rated as less strong than the strong labels (M = 4.25, SD = .82), $t(110) = -3.82$, $p < .001$.

Similarity Judgments

As in Experiments 1 and 2, similarity judgments among the three different dimensions of similarity showed a high positive correlation and thus were averaged to create a single similarity measure for each pair of silhouettes (average within-subject correlations among the three dimensions were r = .63 for Phase 1, and r = .75 for Phase 2). Data reduction procedure followed exactly the one used in Experiments 1 and 2.

The within-subjects design allowed the computation of a 2 x 2 x 2 x 2 ANOVA with two within-subject factors (*Boundary*: within vs. across; *Time*: "Phase 1" vs. "Phase 2") and two between-subject factors *(Labels*: "combined/no label control" vs. "combined label"; *Gender* of the participants). The means and standard errors are presented in Figure 4.

Within-boundary pairs increased in similarity more than did across-boundary pairs, as shown by the interaction between *Time* and *Boundary*, $F(1, 211) = 60.07$, $p < .001$, $\eta^2 = .22$. Most importantly, the expected interaction between *Time*, *Boundary,* and *Labels* was also significant, $F(1, 211) = 6.64$, $p = .01$, $\eta^2 = .03$, indicating stronger categorization effects for labeled than nonlabeled conditions. *Gender* of the participants was significant only as a main effect, $F(1, 211) = 10.00$, $p < .01$, $\eta^2 = .04$, showing that in general female participants perceived the silhouettes to be more

similar than male participants did so this factor was not considered any further in the analyses.

Since the label conditions differ from the control in terms of categorization effects, we now turn to the effects of strength of label and implication. As in Experiment 1 and 2, to sort out the separate effect of within-category similarity (assimilation) and across-category similarity (contrast effect), 2 x 2 x 2 ANOVAs (*Time*: "Phase "1 vs. "Phase 2"; S*trength*: "weak category labels" vs. "strong category label"; *Implications:* descriptive vs. medical) were run on the label conditions only (N = 143) separately on within- and across-boundary pairs.

For within-boundary pairs, the expected 2-way interaction between *Time* and *Strength* was significant, $F(1, 139) = 7.43$, $p = .007$, $\eta^2 = .05$, indicating that strong labels increased within-pair similarity judgments more than did weak labels. The interaction between *Time* and *Implications* was also marginally significant, $F(1, 139) = 3.58$, $p = .07$, $\eta^2 = .02$, suggesting that labels with medical implications had stronger effects than did purely descriptive labels. These two effects were additive only, as there was no significant interaction among *Time*, *Strength*, and *Implication.*

For across-boundary pairs, the interaction between *Time* and *Strength* was also significant, $F(1, 139) = 4.59$, $p = .034$, $\eta^2 = .03$, suggesting that strong labels showed a greater decrease in similarity than did weak labels. The interaction between Time and Implications was not significant.

These results provide clear support for the effect of strength of labels on intracategory assimilation and for intercategory accentuation.

Absolute Estimation of Weight

For absolute estimates of weight, the same data reduction procedures used in Experiments 1 and 2 were followed. Three participants who did not provide weight estimates were dropped from the analyses (remaining $N = 212$).

The design was exactly as the one used for similarity judgments. The means and standard errors are presented in Figure 4. .

The expected 3-way interaction between *Time*, *Boundary*, and *Labels* was significant, $F(1, 208) = 6.73$, $p = .01$, $\eta^2 = .03$. *Gender* of the participants was significant only as a main effect, $F(1, 208) = 11.06$, $p < .01$, $\eta^2 = .05$, showing that female participants perceived the silhouettes to be closer in weight than male participants did so this factor was not considered any further in the analyses.

To sort out whether strength and consequences affect absolute judgments, within the label conditions 2 x 2 x 2 ANOVAs (*Time*: "Phase 1" vs. "Phase 2" x S*trength*: "weak category labels" vs. "strong category label" x *Implication:* descriptive vs. medical) were run separately on within- and across-boundary pairs respectively (remaining $N = 141$).

For within-boundary pairs, there was a significant interaction between *Time* and *Strength*, $F(1, 137) = 3.78$, $p = .05$, $\eta^2 = .03$. For the across-boundary pairs, the interaction between *Time* and *Strength* was also marginally significant, $F(1, 137) = 3.00$, $p = .08$, $\eta^2 = .02$. There was no significant effect of Implications, either alone or in interaction with Time. These results provide clear support for the effect of the strength of labels on intracategory assimilation and marginally on intercategory accentuation.[9]

DISCUSSION

Replicating Experiments 1 and 2, labels affected both similarity judgments and absolute estimates. For similarity judgments there was a strong effect of within-category assimilation, and this effect increased independently with both strength of labels and strength of the consequences associated with the labels. Again, absolute estimates showed evidence of both assimilation and accentuation. When considering the labels condition only, absolute weight estimates showed effects of strength, but not of implication, on both assimilation and accentuation effects. These results replicate the strength effect from Experiment 2. As the effects of implication on similarity judgments were independent of the effects of label strength, the differences between weak and strong labels observed in Experiment 2 are unlikely to be due to an implicit assumption that strong labels were more likely than weak labels to be associated with external consequences.

The original goal of varying the consequences or implications of a labeling system was to determine whether this variable could explain the differences associated with strength of label. Since the effects of strength and consequences were independent, this explanation is weakened. The significant effect of consequences, however, is in itself quite important. It suggests that as the social consequences of a labeling system increase, so do its effects on categorization processes. Those category labels that are tied most strongly to social consequences, for good or ill, and independent of the content of the label, are also likely to show assimilation and accentuation effects. It suggests that the "pragmatic" aspects of classification labels deserve further investigation.

GENERAL RESULTS AND DISCUSSION

The three experiments provide consistent evidence that the judgment of categorized objects is strongly affected by the presence and strength of associated category labels. Participants judged individuals sharing the same label as more similar than those having different labels and this effect increased as the strength of the labels increased. In addition, and perhaps most surprisingly, these effects are independent of the source and consequences of the labels. Experiment 1 showed a labeling effect, even when the source of the label was the participant him- or herself. Participants' ratings of the "strength" of the labels modestly predicted the strength of the categorization effects. Categorization effects were somewhat weaker for self-generated labels than for stronger labels provided in the peer-generated condition, and Experiment 2 was thus designed to clarify the independent influences of category strength and source of the labels. Experiment 2, consistent with Experiment 1, clearly showed that while the strength of the labels was important, the source of the labels (a medical expert or one's college peer) was not. As one possible difference between the weak and strong label condition could be attributed to implied medical treatment in the strong label condition, Experiment 3 was designed to ex-

9. In this experiment, the correlation between the strength index and the dependent measures did not reach standard level of significance probably due to the low variability of the perceived strength of labels (present only in two forms: weak and strong).

plicitly associate weak and strong labels with either clear medical implications or no medical implications, to test whether the effect of labels' strength was independent of a label's possible pragmatic implications. Again, it was the strength of the labels that mattered most, while the implications of the labels increased the effect only slightly and independently of the strength of the labels. Perhaps the simplest summary of these three experiments is that the mere process of labeling a continuum affects the categorization judgment of classified objects, which increases with the strength of the category labels, but is relatively uninfluenced by the source of the category, whether self versus other, or peer versus expert.

In this study, as in all similar research, separating the effect of the labeling process from the putative information conveyed by the labels is an important issue. The experimental findings presented here, especially the strong effects of self-generated labels (Experiment 1), and the negligible effects of expert sources (Experiment 2), suggest that the role of "expert knowledge"—at least in these experiments—is not strong. Experiment 1 showed strong labeling effects even when there was no external source for the labels, since the labels were generated by the subjects themselves. This condition is probably the most convincing evidence for a "mere" labeling process, since participants generated their own labels and, nevertheless, were subsequently affected by them. However, the simple naming an object has been shown to affect memory of that object (Lupyan, 2008a). The simple act of naming a section of a continuum may imply that a category is more richly structured than it actually is (e.g., Hall & Moore, 1997; Markman, 1989).

ASSIMILATION VERSUS ACCENTUATION

Across three studies, labels consistently showed assimilation effects for both measures, whereas accentuation effects were found more consistently for the absolute weight estimates. To clarify some of the apparent inconsistencies in assimilation and contrast effects for the two measures, we aggregated the findings across the three experiments to increase statistical power.[10] Means and standard errors for all participants in the control, weak label and strong label conditions across the three experiments are reported in Figure 5. For similarity judgments, the label conditions, in comparison to the control, show strong assimilation effects, $F(1, 575) = 29.83$, $p < .001$, $\eta^2 = .05$, but somewhat weaker contrast effects, $F(1, 575) = 3.13$, $p = .078$, $\eta^2 = .01$. Planned orthogonal contrasts showed that assimilation effects are present for both weak and strong labels, $F(2, 575) = 20.99$, $p = .01$ and $p < .001$ respectively. Moreover, a contrast effect is also present for strong labels, $F(2, 575) = 2.75$, $p = .022$. For absolute estimates ($N = 573$), both assimilation, $F(1, 571) = 20.51$, $p < .001$, $\eta^2 = .03$ and contrast effects, $F(1, 571) = 26.49$, $p < .001$, $\eta^2 = .04$, showed the expected pattern of results. Planned orthogonal contrasts showed that both weak and strong label conditions differed from controls for both assimilation, $F(2, 571) = 11.10$, $p = .001$ and $p < .001$ respectively and contrast, $F(2, 571) = 15.19$, $p = .001$ and $p < .001$ respectively.

10. Participants from each experiment were used as individual entry ($N = 577$). According to their original condition participants were classified as part of the combined control condition, weak-label condition or strong-label condition. Participants from the self-generated label condition (Exp. 1) were not included ($N = 74$). Statistic analyses were run following the ones used for the Experiments 2 and 3.

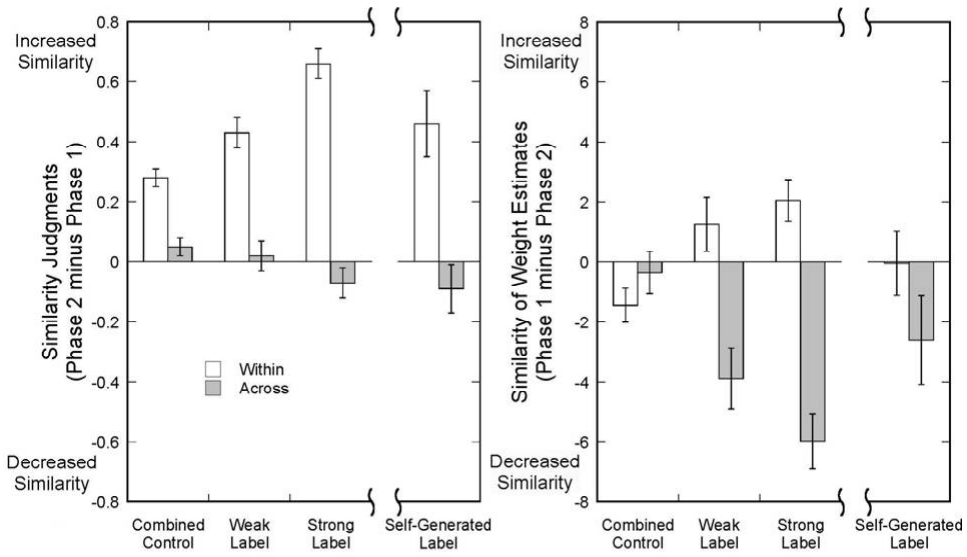

FIGURE 5. Experiment 1, 2, and 3 combined. Difference in similarity judgments (left panel) and weight estimates (right panel) from Phase 1 to Phase 2 for within- and across-boundary pairs by Label condition. Error bars represent standard error of the mean. Positive scores represent increased similarity (i.e., assimilation effects) while negative scores represent decrease similarity (i.e., accentuation effect). Weak labels were: below-average, average, above-average; Strong labels were: anorexic, normal, obese.

The results of this aggregation suggest the presence of both assimilation and contrast effects for both measures. While assimilation seems to be the stronger effect for similarity judgments, both effects are strong for absolute weight estimates. This asymmetry may be at least in part possible because of the types of judgment. Similarity ratings seem especially prone to producing compression effects (Goldstone et al., 2001; Levin & Beale, 2000), and our use of similarity judgments—what Goldstone and colleagues called direct similarity judgments—also produced strong assimilation effects. Our use of absolute weight estimates appears to be closer to Goldstone et al.'s use of a noncategorized standard to avoid experimenter demands. Our original reason for including both direct similarity judgments as well as absolute weight estimates was to compare their susceptibility to categorization effects. We reasoned that similarity judgments, especially unanchored global similarity judgments, might be relatively malleable and easily influenced by category labels, whereas absolute judgments, being anchored to reality and highly familiar, would be harder to influence. It is worth noting again, however, that it was our absolute measures that showed consistently both assimilation and accentuation effects.

Although, the present research was not meant to test alternative accounts of the way categorization affects judgments the results are in line with the idea that category learning and labels—by means of grouping objects together—may alter the representation of the objects (see Goldstone et al., 2001). Verbal labels (particularly when strong labels) may induce categorical representations that reduce differences between members of the same category while exaggerating the differences between members of different categories (Lupyan, 2008b).

For the similarity and absolute measures used in this research, there are important differences in the relation between assimilation and contrast. For direct similarity judgments, participants could show assimilation with limited or no contrast since the sets of within- and across-boundary pairs are independent. On the other hand, for absolute estimates of weight, assimilation and contrast effects are no longer independent. For example, if silhouette 3 is assimilated downward toward the central tendency of its category and silhouette 4 is assimilated upward toward its category's central tendency, then by necessity, accentuation at this boundary (between silhouettes 3 and 4) has been created. In this example, assimilation effects will always result in boundary contrast and vice versa. When the two effects are measure by independent judgments (relative similarity, in this case) then assimilation is the stronger effect (see Goldstone et al., 2001). Such differences between these two types of judgments (relative similarity and absolute estimates) should be considered in planning future research that aims to study separately accentuation and assimilation effect (but see also Corneille, Goldstone, Queller, & Potter, 2006).

ARE CATEGORIZATION EFFECTS INFLUENCED BY THE JUDGED STRENGTH OF CATEGORY LABELS?

The present research shows that the process of labeling a continuum affects the categorization judgment of classified objects, which increases with the strength of the category labels. The set of strong and weak labels, selected based on a pretest, differ on each dimension of category strength: (1) the evaluative differences between adjacent categories, (2) the "categoricalness" of the labels, that is, the degree to which they imply discreteness and discontinuity between adjacent labels, and (3) the impermeability or perceived difficulty of moving from one category to another. At the end of each experiment participants also rated the category labels on the three dimensions. These three ratings were then combined into a single index of perceived category strength that showed to be correlated with two dependent measures. Based on this index, it is possible to determine whether the experimental effects of category strength are dependent upon judged category strength. For the major effects described in the section on the aggregated data (Figure 5), it is possible to examine whether those effects are still significantly present when judged category strength is included as a covariate (testing the effects of labels once the judged strength of the labels is controlled for).

For similarity judgments, the label conditions show strong assimilation effects when judged category strength is not included as a covariate, $F(1, 575) = 29.83$, $p < .001$, $\eta^2 = .05$, but no significant effects when included as a covariate, $F(1, 347) = .71$, *ns*. Contrast effect was only marginal, $F(1, 575) = 3.10$, $p = .078$, $\eta^2 = .01$ and it becomes even weaker and nonsignificant when judged strength is included as a covariate, $F(1, 347) = .14$, *ns*. Planned orthogonal contrasts for similarity judgments showed assimilation effects for both weak and strong labels, $F(2, 575) = 20.99$, $p = .01$ and $p < .001$ respectively, that are not present when judged strength is included as a covariate, $F(2, 346) = 3.82$, *ns.* and *ns.* respectively. A significant contrast effect is also present for strong labels, $F(2, 575) = 2.75$, $p = .022$, but it turns nonsignificant when judged strength is used as a covariate, $F(2, 346) = .58$, *ns*.

For absolute estimates, both assimilation and contrast effects were highly significant, $F(1, 571) = 20.51$, $p < .001$, $\eta^2 = .03$, and, $F(1, 571) = 26.49$, $p < .001$, $\eta^2 = .04$), respectively), but they became much smaller when judged strength is included as a covariate, $F(1, 343) = 5.56$, $p = .02$, $\eta^2 = .01$ and $F(1, 343) = 4.03$, $p = .045$, $\eta^2 = .01$, respectively. Planned orthogonal contrasts for absolute estimates showed both assimilation and contrast effects for both weak and strong labels, for assimilation, $F(2, 571) = 11.10$, $p = .001$ and $p = .001$, and for contrast, $F(2, 571) = 15.19$, $p = .001$ and $p = .001$. However, when judged strength is included as a covariate, assimilation effects are reduced significantly for both weak and strong labels, $F(2, 342) = 3.41$, $p = .05$ and $p = .01$ respectively, and contrast effects are also reduced significantly, $F(2, 342) = 3.04$, *ns*, and $p = .02$ respectively. In summary, the judged strength of category labels is related to the magnitude of categorization effects. It is also clear that when differences in judged strength are controlled for, the labeling effects disappear or are substantially reduced.

WHAT CONSTITUTES A STRONG LABEL?

We argued there are at least three components to the dimension of category strength: (a) evaluative differences between adjacent categories, (b) perceived impermeability of category boundaries, or difficulty in moving from one category to another, and (c) perceived discreteness of the underlying continuum. The weak and strong categories chosen for this research differ on all three dimensions. Labels that differ on these dimensions of strength very likely differ on other dimensions as well. Sloutsky and Fisher (2004) suggested with the relative weight of the visual information in comparison to the weight of the labels, is not fixed. That is,, strong labels, for example, may be more powerful in overwriting visual information in comparison to weak labels. One potentially important aspect of the labeling system is the implicit causal theory that may be activated by the category labels. Causal theories may play an especially important role in defining the meaning of categories (e.g., Ahn, Kim, Lassaline, & Dennis, 2000; Medin & Ortony, 1989). One possible hypothesis is that "strong" categorical labeling systems evoke causal theories that are essentialistic in character. A person who is categorized as "above average" in ponderosity may reasonably conclude that with effort they could move into the average category. However, someone who is classified as "obese" may think of that category as more of a "natural kind" and fixed by one's biological makeup. If so, it is the implicit causal theory associated with the category that generates its strength. This is a topic worthy of further thought and research.

It is also important to note that in the aggregated analysis of Figure 5, the weak label conditions also showed significant assimilation and contrast effects when compared to the control. In our view, this is quite an important result, since the weak labels "below average," "average," "above average" were chosen explicitly to (1) minimize the evaluative differences between categories, (2) emphasize the continuous rather than categorical nature of the underlying metric, and (3) emphasize the ease of movement from one category to another, especially in light of the fuzzy boundaries implied by the category labels. Despite the weakness of these labels, they nonetheless showed significant differences from the unlabeled conditions. In our view, this is further support for the earlier assertion that even the most minimal labeling appears to produce categorization effects.

CONCLUDING REMARKS

The present experiments examined the effects of labels on the judgments of categorized objects using a paradigm that reduced the ambiguity in both the stimulus set and the response scale. The stimuli were simple visual line drawings of body types (silhouettes), familiar to the subjects, and present at the time of judgment. The response scale included two judgments: a judgment of similarity (a relative scale with no familiar anchor points), and a judgment of weight in pounds (an absolute scale, highly familiar to participants and anchored to their real world experiences).

Even though the use of the same stimulus set and paradigm may reduce direct generazability, the present research shows a consistent effect of labels on those judgments. Moreover, the effect of labels in this research design is the result of a within-participant change, since all subjects made their judgments first on an uncategorized continuum. We would expect stronger effects of labeling as the setting becomes more complex, the stimuli more ambiguous, the judgments more difficult or unfamiliar or memory-based, and as time pressure and/or cognitive load increase. These "more complex settings" are nothing other than those that occur in everyday life.

REFERENCES

Ahn, W., Kim, N.S., Lassaline, M.E., & Dennis, M. (2000). Causal status as a determinant of feature centrality. *Cognitive Psychology*, *41*, 361-416.

Allport, G. W. (1954). *The nature of prejudice*. Cambridge, MA: Addison-Wesley.

Campbell, D. T. (1958). Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science*, *3*, 14-25.

Carnaghi, A., Maass, A., Gresta, S., Bianchi, M., Cadinu, M., & Arcuri, L. (2008). Nomina sunt omina: On the inductive potential of nouns and adjectives in person perception. *Journal of Personality and Social Psychology*, *94*, 839-859.

Corneille, O., Goldstone, R. L., Queller, S., & Potter, T. (2006). Asymmetries in the categorization, perceptual discrimination, and visual search for reference and nonreference exemplars. *Memory & Cognition*, *33*, 556-567.

Corneille, O., Klein, O., Lambert, S., & Judd, C. M. (2002). On the role of familiarity with units of measurement in categorical accentuation: Tajfel and Wilkes (1963) revisited and replicated. *Psychological Science, 13*, 380-383.

Eiser, J. R., & Stroebe, W. (1972). *Categorization and social judgment.* New York: Academic Press.

Furnham, A., & Alibhai, N. (1983). Cross-cultural differences in the perception of female body shapes. *Psychological Medicine, 13*, 829-837.

Gelman, S. A., & Heyman, G. D. (1999). Carrot-eaters and creature-believers: The effects of lexicalization on children's inferences about social categories. *Psychological Science, 10*, 489-493.

Goldstone, R. (1995). Effects of categorization on color-perception. *Psychological Science, 6*(5), 298-304.

Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition, 52*, 125-157.

Goldstone, R., Lippa, Y., & Shiffrin, R. (2001). Altering object representations through category learning. *Cognition*, *78*(1), 27-43.

Hall, D. G., & Moore, C. E. (1997). Red bluebirds and black greenflies: Preschoolers' understanding of the semantics of adjec-

tives and count nouns. *Journal of Experimental Child Psychology*, *67*, 236-267.

Harnad, S. (Ed.). (1987). *Categorical perception: The groundwork of cognition*. New York: Cambridge University Press.

Katz, P. A., Karp, B., & Yalisove, D. (1970). Verbal mediation of childrens perception—Role of response variables. *Journal of Experimental Psychology*, *85*, 349-355.

Kobrynowicz, D., & Biernat, M. (1997). Decoding subjective evaluations: How stereotypes provide shifting standards. *Journal of Experimental Social Psychology, 33*, 579-601.

Krueger, J., & Clement, R. W. (1994). Memory-based judgments about multiple categories: A revision and extension of Tajfel's accentuation theory. *Journal of Personality and Social Psychology, 67*, 35-47.

Lakoff, G. (1987). Cognitive model and prototype theory. In U. Nisser (Ed.), *Concept and conceptual development: Ecological and intellectual factors in categorization* (pp. 63-100). New York: Cambridge University Press.

Levin, D. T., & Beale, J. (2000). Categorical perception occurs in newly learned faces, other-race faces, and inverted faces. *Perception and Psychophysics, 23,* 1153-1169.

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology-Learning Memory and Cognition, 24*, 732-753.

Lupyan, G. (2008a). From chair to "chair": A representational shift account of object labeling effects on memory. *Journal of Experimental Psychology: General*, *137*(2), 348-369.

Lupyan, G. (2008b). The conceptual grouping effect: Categories matter (and named categories matter more). *Cognition*, *108*, 566-577.

Markman, E. M. (1989). *Categorization and naming in children*. Cambridge, MA: MIT Press.

McGarty, C., & Turner, J. C. (1992). The effects of categorization on social judgement. *British Journal of Social Psychology, 31*, 253-268.

Medin, D.L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179-195). New York: Cambridge University Press.

Putnam, H. (1975). The meaning of meaning. In H. Putnam (Ed.), *Mind, language, and reality* (Vol. 2). London: Cambridge University Press.

Putnam, H. (1988). *Representation and reality*. Cambridge, MA: MIT Press.

Richardson, G. P. (1991). *Category names and category learning*. Unpublished doctoral dissertation, University of Stirling, Stirling, Scotland.

Robinson, J. S. (1955). The effect of learning verbal labels for stimuli on their later discrimination. *Journal of Experimental Psychology*, *49*, 112-114.

Rothbart, M., Davis-Stitt, C., & Hill, J. (1997). Effects of arbitrarily placed category boundaries on similarity judgments. *Journal of Experimental Social Psychology, 33*, 122-145.

Sloutsky, V., & Fisher, A. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology-General*, *133*(2), 166-188.

Tajfel, H. (1969). Cognitive aspects of prejudice. *Journal of Social Issues, 25*, 79-98.

Tajfel, H., & Wilkes, A. L. (1963). Classification and quantitative judgment. *British Journal of Social Psychology, 54*, 101-114.

Tighe, L. S., & Tighe, T. J. (1966). Discrimination learning: Two views in historical perspective. *Psychological Bulletin, 66,* 353-370.

Walton, G. M., & Banaji, M. R. (2004). Being what you say: The effect of essentialist linguistic labels on preferences. *Social Cognition, 22,* 193-213.

Waxman, S., & Gelman, S. (2009). Early word-learning entails reference, not merely associations. *Trends in Cognitive Sciences*, *13*(6), 258-263.

Waxman, S., & Markow, D. (1995). Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology*, *29*(3), 257-302.