The Effect of Familiarity with the Response Category Labels on Item Response to Likert Scales

BERT WEIJTERS

MAGGIE GEUENS

HANS BAUMGARTNER

Bert Weijters (bert.weijters@ugent.be) is assistant professor of consumer psychology at Vlerick Business School and Ghent University; Dunantlaan 2, 9000 Ghent, Belgium. Maggie Geuens (maggie.geuens@ugent.be) is professor of marketing at Ghent University and Vlerick Business School, Tweekerkenstraat 2, 9000 Ghent, Belgium. Hans Baumgartner (hansbaumgartner@psu.edu) is the Smeal Professor of Marketing in the Smeal College of Business at The Pennsylvania State University, 482 Business Building, University Park, PA 16802.

Correspondence: Bert Weijters. The authors acknowledge the helpful input of the editor, associate editor, and reviewers. In addition, the authors are grateful to the Vlerick Academic Research Fund, partially subsidized by the Flemish government, for their financial support of this research project.

Surveys in the social sciences often employ rating scales anchored by response category labels such as “strongly (dis)agree” or “completely (dis)agree”. Although these labels may exert a systematic influence on responses since they are common to all items, academic research on the effect of different labels is surprisingly scarce. In order to help researchers choose appropriate category labels, we contrast the intensity hypothesis (which posits that response categories are endorsed less frequently if the labels are more extreme) with the familiarity hypothesis (which states that response categories are endorsed more frequently if the labels are more common in day-to-day language). In a series of studies we find consistent support for the familiarity hypothesis. Our results have important implications for the appropriate use of category labels in multilingual surveys and we propose a procedure based on Internet search engine hits to equate labels in different languages in terms of familiarity.

Imagine that a French researcher wants to replicate an empirical finding that was established in the US using data based on consumer self-reports in France. To conduct the study, the researcher will have to translate not only the actual survey items but also the response category labels (e.g., "strongly (dis)agree") by choosing from alternative translations of the labels. For example, while "fortement d'accord" is the literal translation of "strongly agree," the more idiomatic translation "tout à fait d'accord" is also a viable option. Will this seemingly minor variation in the amplifiers used to mark the endpoints of the scale ("fortement" vs. "tout à fait") have an effect on how many respondents endorse the extreme scale categories? Assuming that the category labels influence consumers' responses, what causes this effect? And what are the implications for multilingual and monolingual surveys?

Although a variety of characteristics of rating scales have been studied in prior research (particularly the number of response categories; see Cox 1980), the problem of choosing appropriate labels for the response categories has been largely ignored. This is surprising because category labels typically apply to many if not all of the items in a questionnaire, and they may therefore exert a systematic influence on all responses (in contrast to item wording effects, which are restricted to individual items or subsets of items).

The few studies that have investigated the labels attached to the response alternatives on rating scales suggest that the labels used can affect the response distributions, and researchers have usually attributed differences in responding to the perceived intensity of the labels. That is, respondents are less likely to select a response option if the associated label is more intense (Wildt and Mazis 1978; Wyatt and Meyers 1987). In general, intensity refers to the degree to which an object has the attribute (dimension of judgment) expressed by a label (e.g., excellent vs. good when rating a product's quality), although in this article we are specifically interested in the

extremity or strength of the amplifiers used to anchor the attribute of interest (e.g., complete vs. strong (dis)agreement). We report a detailed investigation of this intensity hypothesis and contrast it with another possible mechanism leading to systematic effects of category labels on response behavior, which we call the familiarity hypothesis. According to the familiarity hypothesis, scale categories marked by labels that are used more often in day-to-day language are more likely to be endorsed in questionnaires. Although both the intensity and familiarity hypotheses are viable mechanisms underlying the effect of category labels on response behavior, our findings provide stronger support for the familiarity hypothesis.

The remainder of this article is structured as follows. After developing the intensity and familiarity hypotheses in detail, we initially test our predictions in a single-language context. This approach helps to safeguard internal validity, and the results are relevant because even when only one language is involved, findings may be compared across studies in which different scale labels were used (e.g., in meta-analytic studies). However, the effect of category labels on responses is particularly relevant in multilingual research when findings are compared across different languages. In this situation, response category labels will necessarily differ across languages because the labels have to be translated. Therefore, once the effect has been established in a monolingual setting, we report two studies that examine the intensity and familiarity hypotheses in a multilingual context.

Our research contributes to the consumer behavior literature in several ways. In the most general sense, we provide evidence that researchers have to pay greater attention to the choice of the category labels used on rating scales (especially the labels employed for the endpoints), because seemingly minor variations in the amplifiers used (e.g., “strongly” vs. “completely” (dis)agree) can influence questionnaire responses. We also show in a head-to-head comparison

of the intensity and familiarity hypothesis that even though the intensity hypothesis is the mechanism discussed in prior literature, differences in response behavior caused by the category labels are more in line with the familiarity hypothesis than with the intensity hypothesis. This implies that when findings are compared across studies, it is important to equate category labels in terms of familiarity. Finally, we demonstrate the relevance of the category labeling effect in multilingual research and advance the novel proposition that researchers conducting surveys in multiple languages have to make sure that the category labels used are equally idiomatic in different languages across which results are to be compared.

CONCEPTUAL DEVELOPMENT

A limited number of studies provide evidence that certain aspects of the labels attached to the response categories on rating scales can have systematic effects on people's responses to questionnaires. For example, the range of response alternatives provided (e.g., relatively low vs. high frequencies of watching TV) can influence respondents' interpretation of, and answer to, questions (Schwarz et al. 1988). Also, the use of different numeric values (-5 to +5 vs. 0 to 10) can change the meaning of endpoint labels such as "not at all successful" (Schwarz et al. 1991). In this article we are specifically interested in whether endpoint labels in Likert scales that differ in terms of the amplifier used (e.g., "strongly" vs. "completely" (dis)agree) can change responses. Some researchers have acknowledged that this might be the case, and the effect has been attributed to differences in the perceived intensity of the amplifiers used. We will first describe this intensity hypothesis and then contrast it with a novel alternative account, which we call the familiarity hypothesis.

The Intensity Hypothesis

It is well-established that the labels assigned to the categories of a rating scale can differ in perceived intensity (i.e., extremity or strength on the underlying dimension of judgment). Several studies have reported scale values of adjectives or adverbs that may be used as scale anchors (e.g., adjectives for evaluating products, such as “good,” “terrific,” or “superior,” as in Wildt and Mazis 1978), and some have specifically investigated the intensity of labels varying in the perceived degree of agreement or disagreement that they imply (e.g., “slightly,” “somewhat” or “very much” agree, as in Spector 1976).

Since more intense labels represent more extreme positions on the psychological continuum of interest and since more extreme positions are typically more exceptional, respondents should select scale categories marked by more intense labels less often. Although differences in perceived label intensity have been demonstrated, surprisingly few studies have investigated the effect of label intensity on questionnaire responses. In some studies, the endpoint labels are held constant and different intermediate labels are studied (Klockars and Yamagishi 1988; Wildt and Mazis 1978), in others the intensity of the endpoint labels is varied (Lam and Stevens 1994; Wyatt and Meyers 1987). In general, label intensity has been shown to influence various aspects of response behavior. For example, Wyatt and Meyers (1987) found that when the extremes of the response scale were anchored by narrower or less absolute labels (i.e., “agree” and “disagree”), responses were distributed more evenly across all five scale steps, whereas when the response scale was bordered by wider or more absolute labels (i.e., “strongly agree” and “strongly disagree”), responses were concentrated more on the intermediate scale steps. Lam and Stevens (1994) showed that this finding depended on the strength of respondents’

opinion about the issue under consideration and the strength of the wording of the items (e.g., “I hate going to school” vs. “I don’t like to go to school”).

Multi-item scales used in consumer research typically use rating scales with endpoint labels consisting of an amplifier (e.g., “strongly,” “completely”) and a verb or adjective (e.g., “agree,” “true”). The question is whether the choice of different amplifiers to mark the endpoints of the scale (e.g., “strongly,” “completely,” or “fully” (dis)agree) affects the extent to which the corresponding category is selected. Researchers typically do not pay much attention to this issue, perhaps assuming that the different options are equivalent. The validity of this assumption may be questioned. Cliff (1959) found that amplifiers such as “decidedly,” “very” or “extremely” multiplied the intensity of the adjectives they modified, and he showed that “extremely” had a stronger intensifying effect than “very,” which in turn had a stronger effect than “decidedly”. Smith et al. (2009) also noted differences in the perceived intensity of the amplifiers investigated in their research (e.g., “completely,” “definitely,” “strongly,” “very much”). However, these authors did not test whether differences in the intensity of the endpoint labels induced differences in the extent to which the associated categories were selected. That is, explicit evidence supporting the intensity hypothesis for different amplifiers in Likert scales is lacking. Although the variation in intensity between the various extent of agreement terms used in the present research is smaller than the differences between the endpoint labels examined in previous studies (e.g., “agree” vs. “strongly agree,” as in Wyatt and Meyers 1987), it is possible that labels differing in the intensity of the amplifier of the endpoint response categories influence response behavior. Hence our intensity hypothesis:

$H_{intensity}$: Endpoint response categories are endorsed less frequently if their labels are more intense.

The Familiarity Hypothesis

If researchers have considered the influence of different category labels on responses to questionnaires at all, they have attributed this effect to differences in perceived intensity, as stated in the first hypothesis. Here, we would like to offer an alternative possibility, one that is, perhaps, more surprising. The basic idea is that more familiar labels are more likely to be endorsed.

The intensity perspective builds on the assumption that amplifiers vary along a single dimension of interest: intensity. Apart from their intensity, amplifiers are treated as interchangeable. In linguistic terms, this perspective is consistent with the open choice model, which states that text is composed of individual items that have to be selected from the lexicon, and the only constraint on the concatenation of words is that the rules of grammar be respected. Sinclair (1987) contrasted the open choice model with the idiom principle, according to which combinations of words are used in conventional patterns. This idea is reflected in the phenomenon of collocation. Collocation refers to the habitual juxtaposition (or co-occurrence) of words in text and speech (Durrant 2008). That is, combinations of words that make up a collocation are found to co-occur more often than would be expected based on their individual frequencies. A classic example of collocation is “strong tea,” which is used more frequently and which sounds more natural to native speakers than the alternative expression “powerful tea”. The latter is grammatically correct, but it is typically judged as unidiomatic (Halliday 1966). Of particular importance for our purpose, there is evidence that certain amplifiers used to express intensity of agreement are more commonly used than others (Kennedy 2003). Altenberg (1991, 133) notes that “of the large repertoire of amplifiers available for expressing a high degree of

intensity, speakers rely on a rather limited set of items, and only a few of these are used with great frequency".

Recent research has shown that formulaic sequences (of which collocations form a subset) are not only used more frequently, but are also processed more quickly by language users (Conklin and Schmitt 2008). That is, there is a positive association between frequency of occurrence in language (speech and/or text), familiarity and processing ease. While this link had been previously established for individual words (Gernsbacher 1984; Oppenheimer 2006), it seems to apply to word combinations as well, with collocations generally being high in frequency of use, familiarity, and ease of processing (Conklin and Schmitt 2008; Durrant 2008).

But how does ease of processing influence survey respondents' endorsement of certain response options? Research on meta-cognitive experiences provides theoretical support for the claim that more familiar, high-frequency labels are more likely to be endorsed. Meta-cognitive experiences, such as the fluency or perceived ease with which new information can be processed, have been shown to serve as a source of information because people automatically draw on these meta-cognitive experiences when forming judgments (see Alter and Oppenheimer 2009 and Schwarz 2004 for reviews). In particular, research indicates that (a) repeated exposure to a stimulus has beneficial effects on processing fluency, (b) repeated and more fluently processed statements are more likely to be rated as true, and (c) stimulus repetition and fluent processing increase liking, preference and confidence judgments (Alter et al. 2007; Fang, Singh, and Ahluwalia 2007; Hawkins and Hoch 1992; Unkelbach 2007; Winkielman et al. 2003). As collocations have been shown to be processed more quickly, familiar (vs. unfamiliar) labels should lead to greater processing fluency and are more likely to be confidently chosen as the true and preferred response option. Hence our familiarity hypothesis:

$H_{Familiarity}$: Endpoint response categories are endorsed more frequently if their labels are more familiar.

PILOT STUDY: SCALING INTENSITY AND FAMILIARITY

Before investigating the intensity and familiarity hypotheses, we conducted a pilot study in order to identify endpoint labels to be used in subsequent testing. Specifically, we attempt to find two labels that imply contradictory responses under the two hypotheses, that is, two labels for which both intensity and familiarity are high for one label and low for the other label. Therefore, when comparing the response distributions for these labels, the endorsement rate for the high-intensity and high-familiarity category label should be relatively low if the intensity hypothesis is true, and it should be relatively high if the familiarity hypothesis is true. In the process of identifying labels for use in the main experiments, we also assess the convergent validity of several methods for scaling label intensity and familiarity, which is the secondary objective of the pilot study.

For scaling labels in terms of intensity, Smith et al. (2009) recommend the use of a direct rating method in which respondents rate the intensity of labels on a numeric scale. To provide evidence of convergent validity, we supplement this direct rating method with a pairwise comparison technique in which respondents compare labels two at a time and for each pair choose the one that they perceive to convey a stronger sense of agreement. The number of times a label has been selected over other labels serves as the measure of perceived intensity.

For measuring label familiarity, we use the following methods: (1) direct ratings of label familiarity on a numeric scale; (2) pairwise comparisons of labels in terms of familiarity; (3) a

lexical decision task; and (4) an observed measure of word frequency based on secondary data. The first two methods are analogous to the intensity ratings. Lexical decision tasks are a standard procedure in fluency research and measure how quickly people classify stimuli as words or non-words: the shorter the response latencies, the greater the processing fluency of the words (Fazio 1990; Meyer and Schvaneveldt 1971). As to the fourth method (word frequency), in many languages, corpora (or collections) of texts are available that can be used to count the occurrence of certain words or phrases. As a proxy for frequency counts derived from formal text corpora, we use the number of hits in an online search engine, which is readily available to researchers and can be used in a standardized way across many languages (as demonstrated later).

Method

We used six different amplifiers in Dutch to examine the intensity and familiarity of response category labels expressing different degrees of agreement. We focused on positive labels only since prior research has shown that modifiers are typically balanced (i.e., "completely agree" and "completely disagree" have reciprocal values; see Smith et al. 2009).

In order to conduct a strong test of convergent validity and to avoid response order effects, we used different samples of Dutch-speaking respondents for the direct rating, pairwise comparison, and lexical decision tasks; the method based on Google hits uses secondary data. A first sample (112 respondents from on online panel, M_{age} = 32.03, SD_{age} = 13.67, 66% female) completed a direct rating task using an 11-point scale. Participants evaluated the same six labels on intensity (0 = neutral; 10 = 100% agreement) and familiarity (0 = "we never use this term in day-to-day language"; 10 = "we very often use this term in day-to-day language"). A second

sample (N = 83 undergraduates, 59% female) evaluated all possible pairwise combinations of the six endpoint labels in terms of their perceived intensity ("Which expression indicates the stronger sense of agreement?") and familiarity ("Which expression is more commonly used in day-to-day language?"). A third sample (125 undergraduates, 57% female) was administered a computer-based category label identification task (the lexical decision task) in a lab. Participants were asked to press a button labeled "end category label" for labels pertaining to an end category and to press a button labeled "not an end category label" for other labels. Response latencies were recorded for the six end category labels shown in table 1, along with five non-end category labels (i.e., enigszins eens [somewhat agree], beetje eens [agree a little], matig eens [moderately agree], deels eens [partly agree], redelijk eens [pretty much agree]). Each label was presented twice on the computer monitor in random order. Finally, we registered the number of Google hits for each label, entering as a search term the precise combination of words in the endpoint label (e.g., "sterk eens") and limiting the search to the language (Dutch) and country (Belgium) under investigation.

Results and Discussion

Based on repeated measures ANOVAs for the first two methods and multilevel analysis (to account for individual differences in response times) for the lexical decision task, we found consistent differences across labels (all p's $< .05$; see table 1).

Insert table 1 about here

A comparison of the means reported in table 1 indicates that the various methods used for assessing the intensity and familiarity of the different endpoint labels have very high convergent validity. For intensity, the correlation of the means obtained from the direct rating and pairwise comparison task is .92. The correlations of the means derived from the four familiarity methods range from .96 to .99 (after reverse-scoring reaction times from the lexical decision task). The results indicate that changing the amplifier in the endpoint label leads to consistent differences in intensity and familiarity across labels, with different measures providing convergent results.

"Sterk" [strongly] emerges as the amplifier with the lowest intensity and familiarity scores, and "volledig" [completely] emerges as the amplifier with the highest intensity and familiarity scores. We therefore select "sterk eens" [strongly agree] and "volledig eens" [completely agree] as the labels used in study 2 to distinguish between the intensity and familiarity hypotheses of how category labels affect response behavior. For readability, we will refer to these conditions as "strongly (dis)agree" and "completely (dis)agree", respectively.

STUDY 1: TESTING THE INTENSITY AND FAMILIARITY HYPOTHESES

Study 1 is a first attempt to investigate the effect of endpoint labels differing in intensity and familiarity on response distributions. Specifically, we contrast people's responses based on two alternative endpoint labels, "completely (dis)agree" and "strongly (dis)agree" (in Dutch). As shown in the pilot study, "completely agree" is high in intensity and familiarity and "strongly agree" is low in both. Hence, respondents should be more likely to endorse "completely agree" if the familiarity hypothesis is true and "strongly agree" if the intensity hypothesis is true.

To measure differences in response distributions as a function of the amplifier used for the endpoint labels, we count how often respondents endorse the endpoint positions on the rating scale across a heterogeneous set of items. The heterogeneity of the items ensures that the findings are not content-specific. In some studies, including the current one, we use heterogeneous items sampled from different unrelated scales (Greenleaf 1992a), in others we use a 16-item scale which was specifically developed for measuring extreme response style (Greenleaf 1992b). In both cases, the items share no substantive content and thus are a "pure" measure of respondents' preference for certain scale positions as a function of the category label used.

Method

We conducted an online survey among Dutch-speaking panel members of an online market research agency (N = 218). The respondents ranged in age from 20 to 65 years (M = 43.2, SD = 11.7), 47% were female, and 58% were highly educated (i.e., had schooling beyond secondary school). We randomly assigned respondents to a questionnaire with different endpoint labels. In one version, "completely" was used as the amplifier (N = 107), in the other "strongly" (N = 111). The intermediate categories ("rather disagree," "neither disagree nor agree," "rather agree") had the same labels in both versions. The questionnaire consisted of (a) 16 heterogeneous items (4 pages with 4 items per page) taken from unrelated scales (e.g., "Air pollution is an important worldwide problem," "I often give compliments to others," "Communication is important in a relation") and (b) pairwise comparisons of the two response category labels ("strongly agree" vs. "completely agree") in terms of intensity and familiarity.

The primary dependent variable is the number of endpoint responses across the 16 heterogeneous items (split-half reliability $r = .67$). The pairwise comparison data served as a manipulation check.

Results and Discussion

"Completely agree" was perceived as more intense and more familiar than "strongly agree" by most respondents (i.e., 78% rated "completely" as more intense and 90% rated it as more familiar; $\chi^2(1) = 68.3$, $p < .05$ and $\chi^2(1) = 138.9$, $p < .05$, respectively). Thus, the label manipulation was successful. The mean number of extreme responses was $M = 3.1$ (SE = .26) for "strongly (dis)agree" and $M = 4.4$ (SE = .33) for "completely (dis)agree," and the difference in means is significant ($t(216) = 3.07$, $p < .05$). This result is inconsistent with the intensity hypothesis, but supports the familiarity hypothesis. That is, respondents are more likely to choose the extremes of the rating scale when the associated category labels are more familiar rather than less intense. This finding was obtained even though the perceived intensity of the labels would have predicted the opposite result.

STUDY 2: THE LABEL FAMILIARITY EFFECT IN MULTILINGUAL RESEARCH

The previous study investigated the relative merits of the intensity and familiarity explanations of the effect of endpoint labels on response distributions in a single-language setting and found support for the familiarity hypothesis. As mentioned before, the familiarity hypothesis can operate within a single language, for example when researchers compare data

based on rating scales anchored by different amplifiers (such as ”strongly (dis)agree” or ”completely (dis)agree”). However, the problem is particularly relevant when surveys are administered in multiple languages. Across languages, amplifiers are different by definition, because response labels have to be translated and several viable translations are usually available. Studies 2 and 3 aim to demonstrate the familiarity effect in such multilingual situations.

The methodological literature suggests that when translating response category labels ensuring equivalent intensity is key (Smith et al. 2009). Harzing (2006, 259), for example, points out that “[e]ven though scale anchors might translate into appropriate local equivalents, the intensity associated with these equivalents might be different from the original language.” Although some researchers have acknowledged the importance of choosing category labels that are equally intense in different languages, which reflects a concern with the implications of the intensity hypothesis for multilingual research, it appears that (1) evidence is lacking to support the claim that the intensity of the specific amplifiers used affects response behavior in multilingual research, and (2) the potential relevance of the familiarity hypothesis has been overlooked. Even when a literal translation results in labels that are equally intense in different languages, the expressions may not be equally familiar. If, as suggested by the familiarity hypothesis, scale categories associated with less familiar labels are less likely to be endorsed, differences in the familiarity of category labels in different languages can have a distorting effect on survey responses. The purpose of the current study is to evaluate this possibility by replicating the label familiarity effect of the previous study and extending it to a multi-language setting.

Method

Data were collected through an online panel provider, yielding a sample of N = 982 consisting of approximately 200 English- or French-speaking respondents in each of five regions (i.e., combinations of a particular language and a particular country): English speakers in the US, English speakers in the UK, English speakers in Canada, French speakers in Canada and French speakers in France. The regional subsamples are similar in demographic makeup (see table 2).

Insert table 2 about here

We administered an online survey in which we experimentally manipulated response category labels for the extreme categories, while keeping the intermediate categories constant within language. Specifically, for the English version of the questionnaire we used the following endpoint category labels: (1) strongly, (2) completely, (3) extremely, (4) definitely, (5) fully, and (6) very much (dis)agree. In the French version we used the following labels for the extreme categories: (1) fortement, (2) complètement, (3) extrêmement, (4) définitivement, (5) entièrement, and (6) tout à fait d'accord (en désaccord). The first five labels in English and French are literal translations of one another, the sixth label in both French (tout à fait) and English (very much) is a language-specific expression that does not have a literal counterpart in the other language.

The questionnaire contained the following parts: (1) the 16 heterogeneous items of Greenleaf's (1992b) scale to measure extreme-response style (for the count of endpoint responses the split-half reliability was r = .67); (2) pairwise intensity and familiarity comparisons of the six labels; and (3) socio-demographics. We computed separate intensity and familiarity scores for the labels for each region (language-country combination) by applying Thurstone's (1928) pairwise comparison scaling method. The advantage of this approach is that it results in

intensity and familiarity scores that are comparable across regions, as Thurstone developed this method in order to obtain absolute scores (based on the assumed normality of the underlying continuum). The resulting intensity and familiarity estimates, as well as the average number of endpoint responses per condition, are presented in table 3.

Insert table 3 about here

To provide convergent validity for the familiarity measure, which is of primary interest in this study, we registered the number of Google hits for each endpoint label for each region (the complete and literal text string, e.g., "strongly agree," limiting the search to the language and country of interest). The correlation between the familiarity measure based on self-reports and the natural logarithm of the number of Google hits was .95 for France, .90 for French-speaking Canada, .88 for English-speaking Canada, .89 for the UK, and .92 for the US. Additionally, for the intensity measures we can compare the scores of four labels in English ("completely," "definitely," "strongly," and "very much" agree) with those reported in Smith et al. (2009, table 1); the four labels have the same rank order in terms of intensity and the correlation between the intensity scores is $r = .93$.

Results and Discussion

We use the aggregated data in Table 3 for a linear regression analysis in which the number of endpoint responses acts as the dependent variable. The independent variables are label intensity and label familiarity (based on Thurstone scaling) and four dummy variables that capture the five regions. Preliminary analyses showed that the standard linear regression model

assumptions were met. The only variable that has significant explanatory power is label familiarity; intensity does not show a significant effect and neither do the region dummies. Label familiarity has a significant positive regression weight (standardized $B = .38$, $p < .05$, $R^2 = .14$), indicating that the number of endpoint responses increases as a function of label familiarity, regardless of country and language.

STUDY 3: EXPLAINING ENDPOINT RESPONDING BASED ON FAMILIARITY IN MULTILINGUAL SECONDARY DATA

In typical multilingual surveys, researchers do not systematically vary the labels of the response categories. Also, they usually do not include explicit measures of label familiarity in the questionnaire. The purpose of this final study is to show that familiarity is a viable antecedent of extreme responding differences between regions in an international survey and to illustrate how to construct and use relative measures of familiarity based on secondary data.

Method

The data are based on a consumer survey among a sample of respondents drawn from the online panel of a European data provider. The survey used a stratified sampling technique where each country was represented by a sample of approximately 1,000 respondents, and respondents were chosen such that, in the case of countries in which multiple languages are spoken, the size of each language group was proportional to the real population distribution. The samples were comparable in terms of age and gender distributions (see table 4). We refer to the combination of a particular language and a particular country as a region (e.g., the French-speaking part of

Switzerland is one region). In total 13,520 respondents from 17 European regions participated in the survey.

Insert table 4 about here

Apart from brand-related questions for a major consumer company that are not relevant to the current purposes, the survey contained the Greenleaf scale (1992b), which was used to assess endpoint responding in the current study (split-half reliability $r = .69$). Items and response category labels were translated and back-translated by a translation agency. For the response scale, we used a fully labeled 7-point Likert format with translated equivalents of the English response category labels of "strongly (dis)agree," "(dis)agree," "slightly (dis)agree," and "neutral" (see table 5).

In the current dataset, as in most secondary international datasets, no direct measures of label familiarity are available. We therefore constructed familiarity measures based on the number of Google hits for response category labels in each language. Using the absolute count of hits for the endpoint labels in each language is not a viable option. The reason is that we expect more hits for languages that are spoken by more people, irrespective of the familiarity of the specific label. For example, a low familiarity label in French is likely to get more hits than a high familiarity label in Dutch, simply because there are more French-speaking Internet users than there are Dutch-speaking Internet users. To circumvent this problem, we constructed a relative measure of familiarity by computing the natural logarithm of the ratio of the number of hits for the endpoint category labels (i.e., the labels for categories one and seven) divided by the number of hits for the '(dis)agree' label without the modifier (i.e., categories two and six). Although in the current data taking the logarithm of the ratio did not substantially affect the results, the

transformation serves to normalize the distribution of the variable. The resulting index takes on a positive value if the endpoint labels get more hits than the labels without the modifier, and it takes on a negative value if the endpoint labels get fewer hits (as should typically be the case).

In parallel with the measure of relative familiarity, we also constructed a measure of relative endorsement rates of the endpoints relative to the adjacent, more moderate categories. For each region, we computed the average response frequency of the endpoints (categories one and seven) and the adjacent categories (categories two and six, which correspond to the (dis)agree labels without a modifier). We then divided the frequency of the endpoint categories by the frequency of the adjacent categories and took the natural logarithm (to normalize the distribution of the variable, although the transformation did not affect the significance of the results in the current data). Here, too, a positive value indicates that the endpoints get more responses than the adjacent categories. Table 5 presents the labels, familiarity indices, and endpoint responding measures for all regions.

Insert table 5 about here

Results and Discussion

To analyze the data, we used a linear regression model. The regression model serves to quantify and test the relation between endpoint responses and label familiarity. The 17 regions served as the unit of analysis, and we regressed the log odds of selecting the endpoints (vs. adjacent categories) in the questionnaire on the log odds representing the relative familiarity of the labels of the endpoints (vs. the labels without modifier, which also correspond to the adjacent category labels) according to the number of Google hits. Preliminary analyses showed that the

standard linear regression model assumptions were met. The regression slope was positive and significant (standardized B = .68, $p < .05$, $R^2 = 46\%$). This means that the number of endpoint responses goes up as the familiarity of the endpoint labels increases.

Prior research has shown that there are important differences in the way respondents from different countries respond to questionnaires (unrelated to item content), which may compromise the validity of cross-country comparisons (see table 1 in Harzing 2006 for a review). Past research has attributed differences in response distributions mainly to nationality and national culture. While our study does not directly test for cultural effects (as the differences in endpoint responding do not relate to culture-related variables), our results suggest that response differences that are sometimes attributed to substantive cultural differences might in some cases be caused by translation artifacts, because response category labels are different in different languages by necessity. Our findings demonstrate that different labels may vary in terms of familiarity, which can result in different response patterns across languages. In particular, if the endpoint label used in a certain language is more familiar than the one used in another language, it is likely that the endpoint will be selected more frequently in the former than in the latter language.

GENERAL DISCUSSION

Overview of the Results

Our results indicate that response category labels that are more familiar lead to higher endorsement frequencies of their associated response categories. We reported several studies,

organized into two major stages. First, we studied different response category labels in a single-language setting, with a focus on endpoint labels of the form “amplifier” plus “(dis)agree” (e.g., “strongly (dis)agree”). We found that response category labels using different amplifiers exhibited systematic variation in terms of intensity and familiarity, and we demonstrated that label familiarity affected endpoint responses: endpoint response categories were endorsed more often if they were marked by more familiar labels (despite their higher intensity in the present case).

After establishing the familiarity effect in a single-language setting, we went on to demonstrate the effect in multilingual surveys. We first showed the familiarity effect of response category labels within and across the French and English languages in Canada, the US, the UK and France, using a quasi-experiment. We then found, in a large international secondary data set, that the familiarity of the endpoint category label (relative to the adjacent category label) could be measured post hoc, using data from an Internet search engine, and we demonstrated that the resulting label familiarity scores showed a substantial and significant association with the likelihood that respondents endorsed the endpoint response categories.

Implications

Overall, our results indicate the importance of the wording of response category labels. The methodological literature has previously emphasized the need to equate the intensity of category labels. In this article we have shown that respondents do not simply scale response categories along an intensity dimension and then map their latent response onto the best-matching category, but that the familiarity of the labels also matters (and may actually be a more

important determinant of response behavior). This effect is especially worrisome in situations in which different response category labels are used, either within the same language (e.g., in meta-analyses, when findings based on different response scales are compared across studies) or across different languages (e.g., in cross-national surveys).

If certain labels attract more responses, this leads to bias. Baumgartner and Steenkamp (2001) discuss how extreme responding biases scale scores. If the modal scale response is above (below) the midpoint, average scores will be inflated (deflated). Differences in extreme responding also affect model estimates in multivariate situations. For example, if two samples of respondents answer the same questions using two different rating scales that lead to different frequencies of endpoint responses, this may result in different observed relationships with other constructs. Following Greenleaf (1992a), imagine a simple model where an attitudinal variable, measured on an agreement rating scale, is positively related to an antecedent which is measured on an objective scale (e.g., age in years) and hence not affected by differences in label familiarity. Compared to respondents in the unfamiliar label condition, younger people with moderately negative and older people with moderately positive “true” attitudes will indicate more extreme negative or positive “observed” attitudes in the familiar label condition (i.e., due to the label familiarity effect, the same “true” attitude is expressed more extremely). This will lead to a stronger relationship (steeper slope) between the attitudinal variable and age in the familiar label condition, compared to the unfamiliar label condition.

To illustrate this spurious moderation effect due to label familiarity, we estimated the relation between an attitudinal item and age, using data from study 1. We chose age because it is a variable that is often used in research and it is measured in a format that is not affected by the response category labeling effect. From the set of heterogeneous items available, we selected the

item that showed the strongest correlation with age: Responses to the statement “I try to avoid food that is high in cholesterol” were positively correlated with age (which makes intuitive sense). We regressed the response to this item on age and tested for the moderating effect of response category label (i.e., the label manipulation ‘strongly (dis)agree’ vs. ‘completely (dis)agree’ used in study 1). Figure 1 shows the regression slope for the two conditions. In line with our expectations, the two slope coefficients are significantly different ($\chi^2(1) = 5.32$, $p < .05$; $B_{completely} = .034$, SE = .006; $B_{strongly} = .015$, SE = .006). “Completely (dis)agree” is higher in familiarity than “strongly (dis)agree” (in Dutch) and attracts more endpoint responses, which explains the stronger slope coefficient for the condition in which “completely (dis)agree” was used as the marker for the endpoint labels.

Insert figure 1 about here

This result illustrates that regression slopes between an observable antecedent such as age and an item response can be significantly and substantially different depending on the familiarity of the labels used for the endpoints. This has potentially detrimental effects in several settings. First, in research that aims to replicate findings that were initially obtained in English in another language, the use of translated labels that are relatively low in familiarity may weaken relationships. Second, in studies based on the same language, different effects may be obtained depending on the labels used. And finally, in comparative studies across multiple language groups, a spurious moderator effect of nationality and/or language may be observed that is in reality a translation artifact.

When constructing scales and translating them, researchers typically devote much effort to the optimal translation of the items, but the response category labels are equally important and should receive more attention than they have been accorded in the past. Failure to do so threatens the comparability of responses across languages, as responses may be biased by the category labels shared by all the items in the scale. In a similar vein, measures need to be validated cross-linguistically rather than – or in addition to – cross-nationally, and it should not be assumed, for example, that different language groups within one country (e.g., German-speaking and French-speaking respondents in Switzerland) show measurement invariance.

Our research leads to practical recommendations regarding translation as well. Based on the current research, we champion the use of response category labels that are equally familiar in different languages. In some cases, literal translations may yield such a desirable result, but this is not necessarily the case. To illustrate, consider some of the data from study 3 (see table 5). Although the German and Dutch labels are literal translations ("vollkommen einverstanden" and "volledig eens," both equivalents of "completely agree"), the German expression is more familiar than the Dutch expression (for each of the relevant regions), thus resulting in more endpoint responses. Back-translation of response category labels does not guarantee equivalence of the labels (Brislin 1970), as back-translations may result in literal but not necessarily idiomatic translations and the familiarity of the labels in different languages may differ.

In emotion assessment, special symbols have been developed (sad, neutral and smiling faces, for example) to avoid translation problems (de Langhe et al. 2011). However, these symbols are typically content-specific (e.g., designed to measure emotional valence) and it is not clear which symbols could be used to express degrees of agreement without the need to explain and label the symbols with words (which would in turn suffer from differences in wording

familiarity). Alternatively, one might consider using colors, but unfortunately even color categorization is language-specific (Athanasopoulos 2011). Eliminating the effect of amplifiers by omitting them altogether (e.g., by using ”disagree” and ”agree” as the endpoints) may result in floor and/or ceiling effects as respondents may be unable to differentiate between strong and moderate degrees of agreement. Researchers may also consider using a common language (e.g., English), but this is an option only if all respondents are equally proficient in the chosen language (Harzing 2006). In sum, the translation problem may often prove inescapable.

Hence, as a key recommendation to international researchers, we advocate a decentering approach to translating questionnaires. Decentering refers to “a translation process in which the source and the target language versions are equally important during the translation procedure” (Brislin 1970, 186). When translating labels, decentering is a must, as frequently used labels in one language (e.g., “tout à fait d’accord” in French) may lack idiomatic translations in other languages. It is crucial to keep in mind future ease of translation when developing scales and when labeling the response categories.

A specific procedure to identify equivalent labels in two languages of interest could be as follows, using as an example a situation in which response category labels for a five-point rating scale are needed for a survey in the UK and France. First, formulate several English-French pairs of endpoint labels, consisting of both literal translations and translations that are similar in terms of familiarity (e.g., the labels used in study 2). Second, look up the number of verbatim search engine hits for each expression in the specific country of interest (e.g., the number of Advanced Google Search hits for the verbatim expression “strongly agree” in English in the UK). Third, to compute comparable scores, divide the count for each endpoint label by the number of Google hits for the label without the modifier (“agree” or “d’accord” in the example), and take the

natural logarithm of the ratio (e.g., ln [$\#hits_{\text{strongly agree}}/\#hits_{\text{agree}}$]). Fourth, compute the discrepancy in familiarity scores between the English and the French labels. Fifth, plot the result as illustrated in figure 2 and select the label pair that has a low discrepancy in familiarity scores across languages.

Insert figure 2 about here

In the example in figure 2, the preferred amplifier pairs for the endpoint labels would be "strongly"/"fortement," "definitely"/"définitivement," or "completely"/"complètement". Note that in this particular case the literal translation of, say, "strongly" in French happens to be an expression that is equally familiar in the two languages. Figure 3 gives an overview of the proposed procedure. Generally speaking, if labels need to be defined in more than two languages, selecting the label with minimal cross-linguistic variation in relative familiarity scores represents the best option.

Insert figure 3 about here

When working with secondary data or in instances where the translation is beyond the control of the researcher, we strongly recommend that researchers examine the familiarity of response category labels post hoc. Specifically, researchers can implement the approach demonstrated in our last study, where Internet search engine results are used as a proxy for the familiarity of the endpoint category label. Any differences in endorsement rates of the extremes

of the rating scale between different languages should be evaluated against differences in familiarity.

Limitations and Future Research

The current research has focused on the endpoint labels only. We have several reasons for narrowing our scope in this way. First, differences in endpoint responding are a common problem in multilingual research (Harzing 2006) and the endpoint labels used are an important determinant of endpoint responding (Arce-Ferrer 2006). Second, researchers often use scales in which only the endpoints are labeled (Weijters, Cabooter, and Schillewaert 2010), so variation in the labels of the endpoints is of greatest practical relevance. Finally, by manipulating the intensity and familiarity of the endpoint labels only, we have a clean dependent variable (i.e., the degree of endorsement of the endpoint scale categories as a function of the labels attached to these categories), which is not confounded by the intensity or familiarity of adjacent category labels. However, further research is needed to evaluate the effects of choosing different amplifiers for the moderate categories (e.g., “somewhat (dis)agree” or “slightly (dis)agree”) and to assess the effect of the endpoint labels on how the moderate categories are interpreted.

Although our findings implicate primarily the familiarity of the response category labels as a driver of non-substantive differences in response distributions across languages, this does not mean that the intensity of the response labels never matters. It is likely that the intensity of the category labels affects response behavior under at least some circumstances, and it is also possible that intensity and familiarity have complementary effects on respondents’ endorsement of the endpoint categories (i.e., the intensity and familiarity hypotheses need not be mutually

contradictory). For example, the recent research of de Langhe et al. (2011) supports the relevance of label intensity in emotion research. In an interesting series of studies, these authors show that when respondents verbalize emotional reactions in a non-native language, they express systematically more intense emotions than when they respond in their native language. De Langhe et al. call this the anchor contraction effect and show that it is due to the lower perceived emotional intensity of rating scale anchors in a non-native language compared to one's native language (which is related to the extent to which emotional words are part of more episodic traces in memory; Puntoni, De Langhe, and Van Osselaer 2009). Although de Langhe et al. investigate a somewhat different set of issues (they are specifically concerned with differences in responding in one's native vs. another language, and they show that the anchor contraction effect only applies to emotional reactions), it would be interesting to investigate under what conditions the intensity or familiarity of category labels affects response behavior.

Future research is also needed to validate our findings in other languages, including non-European languages. In addition, it would be helpful if a set of response category labels and their corresponding familiarity scores were available in different languages. As a first step, table 3 provides the familiarity scores for the response category labels used in study 2. To complement this limited list and as an easily accessible alternative database, we recommend the use of Internet search engines. Study 3 demonstrated their effectiveness in providing proxy measures of label familiarity, and the approach illustrated in figures 2 and 3 can be used as a method for selecting labels in multilingual survey research.

REFERENCES

Altenberg, Bengt (1991), "Amplifier Collocations in Spoken English," in *English Computer Corpora: Selected Papers and Research Guide*, ed. Stig Johansson and Anna-Brita Stenström, Berlin, New York: Mouton de Gruyter, 127-47.

Alter, Adam L. and Daniel M. Oppenheimer (2009), "Uniting the Tribes of Fluency to Form a Metacognitive Nation," *Personality and Social Psychology Review*, 13 (August), 219-35.

Alter, Adam L., Daniel M. Oppenheimer, Nicholas Epley, and Rebecca N. Eyre (2007), "Overcoming Intuition: Metacognitive Difficulty Activates Analytic Reasoning," *Journal of Experimental Psychology: General*, 136 (November), 569-76.

Arce-Ferrer, Alvara J. (2006), "An Investigation into the Factors Influencing Extreme-Response Style," *Educational and Psychological Measurement*, 66 (June), 374-92.

Athanasopoulos, Panos (2011), "Color and Bilingual Cognition," in *Language and Bilingual Cognition*, ed. Vivian Cook and Benedetta Bassetti, New York: Psychology Press, 241-62.

Baumgartner, Hans and Steenkamp, Jan-Benedict E.M. (2001), "Response Styles in Marketing Research: A Cross-national Investigation," *Journal of Marketing Research*, 38 (May), 143-56.

Brislin, Richard (1970), "Back-translation for Cross-cultural Research," *Journal of Cross-Cultural Psychology*, 1 (September), 185-216.

Cliff, Norman (1959), "Adverbs as Multipliers," *Psychological Review*, 66 (January), 27–44.

Conklin, Kathy and Norbert Schmitt (2008), "Formulaic Sequences: Are They Processed More Quickly than Nonformulaic Language by Native and Nonnative Speakers?" *Applied*

Linguistics, 29 (March), 72-89.

Cox III, Eli P. (1980), "The Optimal Number of Response Alternatives for a Scale: A Review," *Journal of Marketing Research*, 17 (November), 407-22.

de Langhe, Bart, Stefano Puntoni, Daniel Fernandes, and Stijn M.J. van Osselaer (2011), "The Anchor Contraction Effect in International Marketing Research," *Journal of Marketing Research*, 48 (April), 366-80.

Durrant, Philip (2008), *High Frequency Collocations and Second Language Learning*. Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy (unpublished).

Fang, Xiang, Surendra Singh, and Rohini Ahluwalia (2007), "An Examination of Different Explanations for the Mere Exposure Effect," *Journal of Consumer Research*, 34 (June), 97-103.

Fazio, Russell H. (1990), "A Practical Guide to the Use of Response Latency in Social Psychological Research," in *Research Methods in Personality and Social Psychology*, Vol. 11, ed. Clyde Hendrick and Margaret S. Clark, Newbury, CA: Sage, 74-97.

Gernsbacher, Morton Ann (1984), "Resolving 20 Years of Inconsistent Interactions Between Lexical Familiarity and Orthography, Concreteness, and Polysemy," *Journal of Experimental Psychology: General*, 113 (June), 256-81.

Greenleaf, Eric A. (1992a), "Improving Rating Scale Measures by Detecting and Correcting Bias Components in Some Response Styles," *Journal of Marketing Research*, 29 (May), 176–88.

_____ (1992b), "Measuring Extreme Response Style," *Public Opinion Quarterly*, 56 (Autumn), 328-50.

Halliday, Michael A. K. (1966), "Lexis as a Linguistic Level," in *In Memory of J.R. Firth*, ed. Charles E. Bazell, John C. Catford, Michael A. K. Halliday and Robin H. Robins, London: Longmans, Green and Co. Ltd., 148-62.

Harzing, Anne-Wil (2006), "Response Styles in Cross-national Survey Research: A 26-Country Study," *International Journal of Cross-Cultural Management*, 6 (August), 243-66.

Hawkins Scott A. and Stephen J. Hoch (1992), "Low-involvement Learning – Memory Without Evaluation," *Journal of Consumer Research*, 19 (June), 212-25.

Kennedy, Graeme (2003), "Amplifier Collocations in the British National Corpus: Implications for English LanguageTeaching," *Teachers of English to Speakers of Other Languages (TESOL) Quarterly*, 37 (Autumn), 467-87.

Klockars, Alan J. and Midori Yamagishi (1988), "The Influence of Labels and Positions in Rating Scales," *Journal of Educational Measurement*, 25 (Summer), 85-96.

Lam, Tony C.M. and Joseph J. Stevens (1994), "Effects of Content Polarization, Item Wording and Rating Scale Width on Rating Response," *Applied Measurement in Education,* 7 (2), 141-59.

Meyer, David E. and Schvaneveldt, Roger W. (1971), "Facilitation in Recognizing Pairs of Words: Evidence of a Dependence Between Retrieval Operations," *Journal of Experimental Psychology*, 90 (October), 227-34.

Oppenheimer, Daniel M. (2006), "Consequences of Erudite Vernacular Utilized Irrespective of Necessity: Problems with Using Long Words Needlessly," *Applied Cognitive Psychology*, 20 (March), 139-56.

Puntoni, Stefano, Bart De Langhe, and Stijn M.J. Van Osselaer (2009), "Bilingualism and the Emotional Intensity of Advertising Language," *Journal of Consumer Research*, 35

(April), 1013-26.

Schwarz, Norbert (2004), "Metacognitive Experiences in Consumer Judgment and Decision Making," *Journal of Consumer Psychology*, 14 (4), 332-48.

Schwarz, Norbert, Bärbel Knäuper, Hans-J. Hippler, Elisabeth Noelle-Neumann, and Leslie Clark (1991), "Rating Scales: Numeric Values May Change the Meaning of Scale Labels," *Public Opinion Quarterly*, 55 (Winter), 570-82.

Schwarz, Norbert, Fritz Strack, Gesine Müller, and Brigitte Chassein (1988), "The Range of Response Alternatives May Determine the Meaning of the Question: Further Evidence on Informative Functions of Response Alternatives," *Social Cognition*, 6 (2), 107-17.

Sinclair, John M. (1987), "Collocation: A Progress Report," in *Language Topics: Essays in Honour of Michael Halliday*, Vol. 2, ed. Ross Steele and Terry Threadgold, Amsterdam/Philadelphia: John Benjamins Publishing Company, 319-31.

Smith, Tom W., Peter Ph. Mohler, Janet Harkness, and Noriko Onodera (2009), "Methods for Assessing and Calibrating Response Scales across Countries and Languages," in *New Frontiers in Comparative Sociology*, ed. Masamichi Sasaki, Leiden, Netherlands: Brill, 45-79.

Spector, Paul E. (1976), "Choosing Response Categories for Summated Rating Scales," *Journal of Applied Psychology*, 61 (June), 374-75.

Thurstone, Louis Leon (1928), "Attitudes Can be Measured," *American Journal of Sociology*, 33 (January), 529-54.

Unkelbach, Christian (2007), "Reversing the Truth Effect: Learning the Interpretation of Processing Fluency in Judgments of Truth," *Journal of Experimental Psychology: Learning, Memory and Cognition*, 33 (January), 219-30.

Weijters, Bert, Elke Cabooter, and Niels Schillewaert (2010), "The Effect of Rating Scale Format on Response Styles: The Number of Response Categories and Response Category Labels," *International Journal of Research in Marketing*, 27 (September), 236 -47.

Wildt, Albert R. and Michael B. Mazis (1978), "Determinants of Scale Response: Label versus Position," *Journal of Marketing Research, 15* (May), 261–67.

Winkielman, Piotr, Norbert Schwarz, Tetra Fazendeiro, and Rolf Reber (2003), "The Hedonic Marking of Processing Fluency: Implications for Evaluative Judgment," in *The Psychology of Evaluation: Affective Processes in Cognition and Emotion*, ed. Jochen Musch and Karl C. Klauer, Mahwah, NJ: Erlbaum, 189–217.

Wyatt, Randall C. and Lawrence S. Meyers (1987), "Psychometric Properties of 4 5-Point Likert-Type Response Scales," *Educational and Psychological Measurement*, 47 (March), 27-35.

Table 1:

Mean Intensity and familiarity scores in pilot study (with standard errors in parentheses)

	Intensity		Familiarity				Summary measures[e]	
Dutch label (free translation)	Direct Rating[a]	Paired comparison[b]	Direct Rating[a]	Paired comparison[b]	Response latency[c]	Google hits	Intensity	Familiarity
Sterk eens (strongly agree)	7.89 (.15)[1]	.94 (.13)[1]	3.62 (.27)[2]	1.14 (.11)[1]	1011.53 (55.01)[1]	510	- .99	- .74
Zeer eens (very much agree)	7.60 (.18)[1]	1.43 (.12)[2]	2.49 (.24)[1]	1.65 (.12)[2]	1002.93 (54.86)[1]	755	- .95	- .74
Zeker eens (certainly agree)	7.78 (.20)[1]	2.11 (.11)[3]	5.05 (.28)[3]	2.40 (.10)[3]	989.72 (54.61)[1]	n.a.[d]	- .56	- .22
Uitgesproken eens (distinctly agree)	8.57 (.22)[2]	2.98 (.18)[4]	2.81 (.28)[1]	1.18 (.13)[1]	1021.87 (54.73)[1]	55	.25	- .83
Helemaal eens (fully agree)	9.54 (.12)[3]	3.72 (.13)[5]	8.62 (.16)[4]	4.24 (.08)[4]	724.88 (55.83[2]	131000	1.10	1.23
Volledig eens (completely agree)	9.56 (.10)[3]	3.82 (.12)[5]	8.59 (.18)[4]	4.39 (.08)[4]	672.73 (55.21)[2]	110000	1.16	1.25

[a] Sample 1; cell entries represent the mean intensity/familiarity rating (on an 11-point scale).
[b] Sample 2; cell entries represent the mean of the number of times a particular label was judged to be more intense/familiar than the other labels.
[c] Sample 3; cell entries represent the mean response time in milliseconds; smaller numbers indicate a faster response and hence higher familiarity.
[d] Most hits for "zeker eens" refer to the homonym "definitely once" so we treat this value as missing; to compute the correlations between measures, we impute the mean.
[e] The last two columns display the means of the two intensity measures and the four familiarity measures, respectively, based on standardized scores (obtained by subtracting the mean and dividing by the standard deviation of each measure across labels).
[1234] Identical numerical superscripts within a column signify that these labels do not differ significantly from each other; different superscripts imply a significant difference between the labels.

Table 2:

Sample descriptive statistics (study 2)

Country, language	N	% Female	M_{age}	SD_{age}
France, French	226	53.5%	41.3	12.8
Canada, French	199	49.7%	40.5	12.3
Canada, English	193	48.2%	40.8	12.3
UK, English	182	49.5%	41.4	13.1
US, English	182	49.5%	41.5	13.0
Total	982	50.2%	41.1	12.7

Table 3:

Label intensity and familiarity scores and average number of endpoint responses per condition (Study 2)

Language, country	Response category label	Intensity	Familiarity	$M_{endpoint}$
French, France	Fortement d'accord	.31	.15	2.03
	Complètement d'accord	.93	1.07	3.28
	Extrêmement d'accord	.72	.00	2.55
	Définitivement d'accord	.90	.25	2.95
	Entièrement d'accord	1.05	1.25	3.15
	Tout à fait d'accord	.72	1.47	3.39
French, Canada	Fortement d'accord	.76	.00	2.60
	Complètement d'accord	1.15	.70	3.53
	Extrêmement d'accord	1.27	.12	1.91
	Définitivement d'accord	1.19	.28	2.60
	Entièrement d'accord	1.32	.70	2.94
	Tout à fait d'accord	.82	.97	2.35
English, Canada	Strongly agree	.60	.62	2.78
	Completely agree	.82	1.01	2.16
	Extremely agree	.66	.00	2.29
	Definitely agree	.69	.95	2.66
	Fully agree	.58	.73	2.70
	Very much agree	.19	.23	2.97
English, UK	Strongly agree	.82	.85	2.22
	Completely agree	1.10	1.24	2.96
	Extremely agree	.54	.00	3.07
	Definitely agree	.86	.95	2.66
	Fully agree	.82	1.00	2.73
	Very much agree	.24	.32	2.69
English, US	Strongly agree	.72	.63	2.12
	Completely agree	.90	.98	3.70
	Extremely agree	.72	.00	2.37
	Definitely agree	.68	.79	4.00
	Fully agree	.68	.61	2.59
	Very much agree	.25	.14	2.83

Table 4:

Sample descriptive statistics Pan-European study (study 3)

	N	Female	M_{age}	SD_{age}
Dutch, Belgium	644	51%	41.0	11.1
Dutch, Netherlands	1046	50%	40.8	11.4
English, UK	908	56%	41.8	11.3
French, Belgium	371	51%	40.5	11.7
French, France	1000	51%	39.4	11.9
French, Switzerland	303	51%	42.5	9.7
German, Germany	993	50%	39.3	11.0
German, Switzerland	606	48%	43.5	9.4
Hungarian, Hungary	1003	51%	38.3	11.8
Italian, Italy	939	50%	39.0	10.6
Italian, Switzerland	50	56%	32.9	8.7
Polish, Poland	802	37%	32.2	11.0
Romanian, Romania	970	50%	37.9	11.5
Slovakian, Slovakia	1063	50%	38.2	12.1
Spanish, Spain	934	50%	37.8	10.5
Swedish, Sweden	974	49%	39.9	11.3
Turkish, Turkey	914	43%	32.5	9.4
Total	13520	49%	38.7	11.4

Table 5

Response category labels, familiarity measures, and response frequencies by region (study 3)

	Labels					
	1	2	6	7	Endpoint familiarity [a]	Endpoint use [b]
Dutch, Belgium	Volledig oneens	Oneens	Eens	Volledig eens	-7.13	- .82
Dutch, Netherlands	Volledig oneens	Oneens	Eens	Volledig eens	-6.23	-1.02
English, UK	Strongly disagree	Disagree	Agree	Strongly agree	-5.13	- .83
French, Belgium	Pas du tout d'accord	Pas d'accord	D'accord	Tout à fait d'accord	- .38	- .39
French, France	Pas du tout d'accord	Pas d'accord	D'accord	Tout à fait d'accord	- .34	- .14
French, Switzerland	Pas du tout d'accord	Pas d'accord	D'accord	Tout à fait d'accord	-1.00	- .27
German, Germany	Überhaupt nicht einverstanden	Nicht einverstanden	Einverstanden	Vollkommen einverstanden	-2.11	- .49
German, Switzerland	Überhaupt nicht einverstanden	Nicht einverstanden	Einverstanden	Vollkommen einverstanden	-2.29	- .51
Hungarian, Hungary	Egyáltalán nem értek egyet	Nem értek egyet	Egyetértek	Teljesen egyetértek	- .75	- .51
Italian, Italy	Non sono assolutamente d'accordo	Non sono d'accordo	Sono d'accordo	Sono assolutamente d'accordo	-1.19	- .60
Italian, Switzerland	Non sono assolutamente d'accordo	Non sono d'accordo	Sono d'accordo	Sono assolutamente d'accordo	-1.22	- .51
Polish, Poland	Zdecydowanie nie zgadzam się	Nie zgadzam się	Zgadzam się	Zdecydowanie zgadzam się	-5.15	- .38
Romanian, Romania	Nu sunt deloc de acord	Nu sunt de acord	Sunt de acord	Sunt complet de acord	-5.16	- .29
Slovak, Slovakia	Veľmi nesúhlasím	Nesúhlasím	Súhlasím	Veľmi súhlasím	-7.16	-1.22
Spanish, Spain	Muy en desacuerdo	En desacuerdo	De acuerdo	Muy de acuerdo	-3.02	- .69
Swedish, Sweden	Instämmer inte alls	Instämmer inte	Instämmer	Instämmer helt	-1.85	- .55
Turkish, Turkey	Kesinlikle katılmıyorum	Katılmıyorum	Katılıyorum	Kesinlikle katılıyorum	-2.91	- .30

[a] Endpoint familiarity: relative measure of familiarity obtained by computing the natural logarithm of the ratio of the number of Google hits for the endpoint categories (i.e., categories one and seven) divided by the number of hits for the adjacent more moderate categories (i.e., categories two and six). The resulting ratio takes on a positive value if the endpoint labels get more hits than the adjacent moderate labels, and it takes on a negative value if the endpoint labels get fewer hits (as will typically be the case).

[b] Measure of relative endorsement rates of the endpoints relative to the adjacent moderate categories. For each region, we computed the average response frequency of the endpoints (i.e., categories one and seven) and the adjacent more moderate categories (i.e., categories two and six). We then divided the frequency of the endpoint categories by the frequency of the adjacent categories and took the natural logarithm. Negative values indicate that the endpoints attract fewer responses than the adjacent categories.

FIGURE LEGEND PAGE

FIGURE 1

THE MODERATING EFFECT OF RESPONSE CATEGORY LABELS

NOTE. —The figure shows the regression-based predicted item scores given age and response category label condition.

FIGURE 2

PLOT OF ENDPOINT LABEL FAMILIARITY IN TWO LANGUAGES

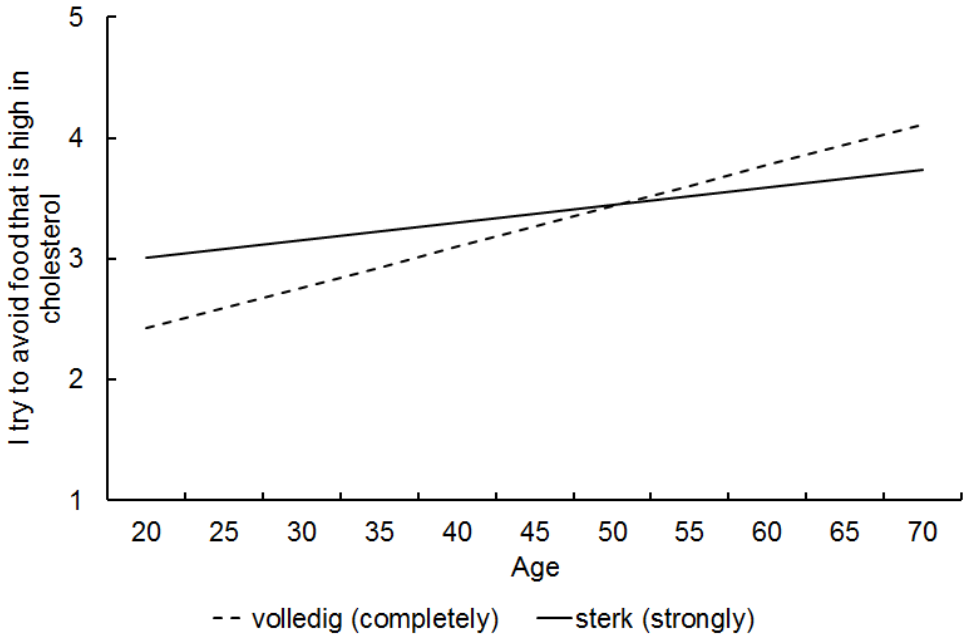
NOTE. —The lines display the familiarity scores of the endpoint labels in English (solid line) and French (dashed line); the bars indicate the discrepancy (absolute difference) in familiarity scores between the English and the French label. Ideal label pairs have a low discrepancy combined with high familiarity in either language. In the graph, “strongly”/“fortement,” “definitely”/ “définitivement,” and “completely”/“complètement” are good candidates as they do well on both counts.

FIGURE 3

A PROPOSED PROCEDURE FOR IDENTIFYING TRANSLATED ENDPOINT LABELS IN TWO LANGUAGES

FIGURE 1:

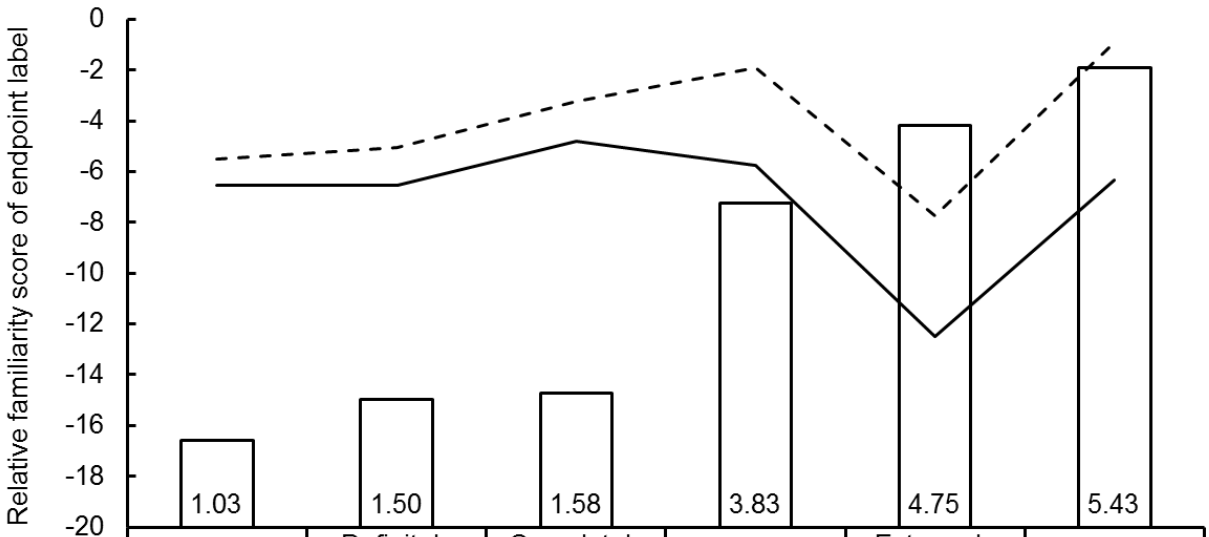
THE MODERATING EFFECT OF RESPONSE CATEGORY LABELS



Note: The figure shows the regression-based predicted item scores given age and response category label condition.

FIGURE 2:

PLOT OF ENDPOINT LABEL FAMILIARITY IN TWO LANGUAGES

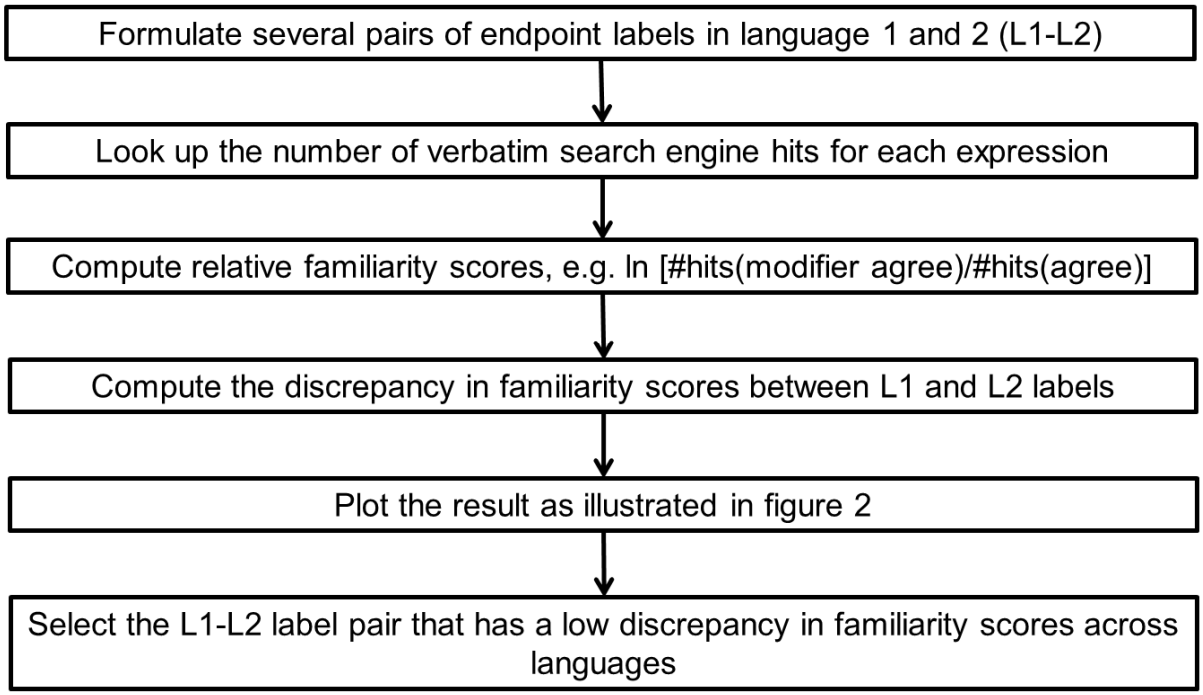


	Strongly agree / Fortement d'accord	Definitely agree / Définitive-ment d'accord	Completely agree / Complète-ment d'accord	Fully agree / Entièrement d'accord	Extremely agree / Extrême-ment d'accord	Very much agree / Tout à fait d'accord
D	1.03	1.50	1.58	3.83	4.75	5.43
ENG	-6.54	-6.55	-4.81	-5.73	-12.50	-6.34
FR	-5.52	-5.05	-3.23	-1.90	-7.75	-0.92

Note: The lines display the familiarity scores of the endpoint labels in English (solid line) and French (dashed line); the bars indicate the discrepancy (absolute difference) in familiarity scores between the English and the French label. Ideal label pairs have a low discrepancy combined with high familiarity in either language. In the graph,"strongly"/"fortement," "definitely"/ "définitivement," and "completely"/"complètement" are good candidates as they do well on both counts.

FIGURE 3:

A PROPOSED PROCEDURE FOR IDENTIFYING TRANSLATED ENDPOINT LABELS IN TWO LANGUAGES



HEADINGS LIST