Published in final edited form as:
Food Qual Prefer. 2013 April 1; 28(1): 36–44. doi:10.1016/j.foodqual.2012.07.012.

Direct comparison of the generalized Visual Analog Scale (gVAS) and general Labeled Magnitude Scale (gLMS)

John E. Hayes[1,2,*], **Alissa L. Allen**[1,2], and **Samantha M. Bennett**[1,2]

[1]Sensory Evaluation Center, The Pennsylvania State University University Park, Pennsylvania

[2]Department of Food Science College of Agricultural Sciences, The Pennsylvania State University University Park, Pennsylvania

Abstract

Hundreds of studies have used the generalized Labeled Magnitude Scale (gLMS) to collect intensity data. Recent work on generalized affective scales like the Labeled Affective Magnitude (LAM) scale and Labeled Hedonic Scale (LHS) suggest a substantial proportion of participants fail to use the entire range of generalized scales, marking only at the adjective labels. This categorical behavior (i.e., clustering) is not limited to affective ratings, as it is well known anecdotally among users of the gLMS. One way to stop this behavior would be to retain a generalized top anchor and cross modal orientation procedure while stripping away the internal adjectives. Several published studies have already used this variant, the generalized Visual Analog Scale (gVAS). Because there are no reports directly comparing the gVAS and gLMS head to head, we did so in two experiments. In Experiment 1, participants (n=87) were randomized to 1 of 3 conditions to test effects of scaling instructions and scale structure. In Experiment 2, participants (n=58) assessed perceived ease of use and resolving power for each scale in a two-session crossover design. gLMS data showed evidence of categorical behavior, while gVAS data did not. Explicitly instructing participants to rate between adjectives did not reduce this behavior. The gLMS was easier to use according to participants, but resulted in non-normal data due to clustering near the adjective labels. gVAS data did not show categorical behavior, as there are no adjectives to cluster around, but the gVAS sacrifices semantic information about the magnitude of response. Regardless of scale type, participants felt the cross-modal orientation procedure helped them understand how to use the scale. Both scales were able to discriminate between sucrose samples in a concentration series. Relative tradeoffs between the two methods suggest the choice of one scale over the other depends on the specific goals and context of the project.

1. Introduction

Humans cannot share perceptual experiences – we can only describe those experiences. To compare experiences to those experienced by others, we first need to transform our internal experience into a verbal description or number we can share. Category scales and visual analog scales have been used historically to quantify sensory or hedonic experiences (reviewed by Bartoshuk, et al., 2003 and Lim, Wood, & Green, 2009). Category scales partition intensity into bins that frequently have a numeric and/or semantic label (e.g.,

[*]Correspondence to be sent to: John E. Hayes, Ph.D., Assistant Professor of Food Science, Pennsylvania State University, 220 Food Science Building, University Park, PA 16802, 814-863-7129 (voice), jeh40@psu.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1=very weak, 5=medium, 9=very strong) whereas visual analog scales have no subdivisions. Rather, a visual analog scale is typically an unstructured line scale anchored at its ends with the minimum and maximum ratings for a particular attribute (e.g., 'not sweet' to 'extremely sweet'). These scales are purportedly straightforward and easily understood by participants across diverse educational attainment or disease states (e.g., Hayes & Patterson, 1921; Zealley & Aitken, 1969), unlike magnitude estimation, which asks participants to express intensities in terms of ratios, requiring both training and a certain level of numeracy (Lawless & Malone, 1986a). Line and category scales are also faster to use, and easier to understand than magnitude estimation (Lawless & Malone, 1986b). Thus, the Labeled Magnitude Scale – a specialized line scale with semantic labels at empirically derived intervals – was rapidly adopted in chemosensory research, as it generates data similar to magnitude estimation (Green, et al., 1996; Green, Shaffer, & Gilmore, 1993) and is easier to use. Unlike visual analog scales and categorical scales, it is assumed magnitude estimation and the Labeled Magnitude Scale (LMS) generate ratio level data.

More critically, visual analog scales, category scales, and the original version of the LMS assume that the adjective labels mean the same thing to all participants within a specific sensory modality; that is, if we are each given solutions of 1 molar sucrose, these methods assume that your 'very sweet' is equivalent to my 'very sweet'. However, work by Bartoshuk and colleagues (2003; 2006) indicates this is not a valid assumption, as it still does not tell us about the perceived intensity of the experience, only that people use the same label or number for that particular experience in that context. In an early description of the visual analog scale, Aitkin (1969) presciently cautioned "the same word can be used with different meaning, and need not imply that people experience the same feeling."

The generally accepted remedy to this problem is two fold: detailed instructions on how to use the scale and an orientation (warmup) procedure to practice using the scale. Initially, Bartoshuk, Green and colleagues suggested that the top of the scale should be anchored to the 'strongest imaginable sensation of any kind' (see Bartoshuk, 2000), resulting in the widespread adoption of the generalized Labeled Magnitude Scale (gLMS) within the chemosensory community (eg, Eggleston, White, & Sheehe, 2010; Hayes & Duffy, 2007; Keast & Roper, 2007). However, subsequent work by Bartoshuk's laboratory indicated the modifier 'imaginable' does *not* generalize the scale, as adding *imaginable* merely stretches the range of values obtained (Snyder, Fast, & Bartoshuk, 2004). Instead, the first remedy is to specifically direct participants to make their ratings in a context broader than (outside) the specific modality being studied; that is, to anchor the top of the scale to the strongest sensation *of any kind*. The second remedy is to have participants practice rating sensation intensities that are not within the modality of interest. Thus, in a hypothetical study of sucrose sweetness, participants would be oriented to the scale by having them rate the intensity of tones or remembered sensations of sound, brightness and temperature (eg, Bartoshuk, et al., 2003; Green & Hayes, 2003). As a secondary benefit, similar to magnitude matching (Marks, et al., 1988), the inclusion of non-oral standards within a test session, such as a remembered sensation like the brightness of the sun or a presented sound standard, allows for data standardization (Duffy, Peterson, & Bartoshuk, 2004) or statistical partitioning of scale usage (Hayes & Duffy, 2007) if needed. With cross modal data, it becomes possible to check whether apparent individual differences in chemosensory are merely the result how that participant uses the scale usage. For example, if two individuals rate the sweetness of a sucrose sample as 20 and 30 on the gLMS, but both rate the brightness of the sun near 60, we are more confident the two individuals perceive the sucrose differently.

In using the gLMS over the last decade, we have observed some participants treat the gLMS labels as categories, marking only at the adjectives, or on a pencil and paper ballots, even circling the words themselves. Similar clustering effects (ie, categorical behavior) in direct scaling are not new; they have been described previously for Magnitude Estimation (Moskowitz, 1977), the Labeled Affective Magnitude (LAM) scale, both with paper (Cardello, Lawless, & Schutz, 2008; Lawless, Sinopoli, & Chapman, 2010) and computerized ballots (Lawless, Popper, & Kroll, 2010) and the Labeled Hedonic Scale (LHS) (Lim & Fujimaru, 2010). While this response pattern is well known anecdotally among practitioners who use the gLMS, there are no formal reports in the literature regarding this response pattern.

One means to prevent categorical behavior would be to strip the gLMS of internal adjectives (the semantic labels), while retaining the top and bottom anchor. This variant, called the generalized Visual Analog Scale (gVAS), is described elsewhere (Dionne, et al., 2005; Snyder, et al., 2006). It has already been used in a large epidemiological trial to collect taste phenotypes in children (Timpson, et al., 2007) and to characterize alcohol beverages in adults (Pickering, et al., 2010). Although the logic behind the gVAS is straightforward (see Dionne, et al., 2005; Snyder, et al., 2006), and it is already being used in the field (see above), there are no reports directly comparing the unstructured gVAS to the gLMS. In pilot analysis comparing the two methods, we found the rank order of imagined sensations from the orientation procedure was preserved, and the correlation across group means was very high. However, we also observed that there was systematic deviation in the raw values obtained from the two scales. Additionally, while categorical behavior was clearly present in the gLMS data, an anonymous reviewer was concerned this response pattern may have arisen due to abbreviated instructions that did not explicitly stress that participants should use the space between the semantic labels (personal communication; anonymous reviewer). Here, we report the results of two new experiments specifically designed to address these questions.

Here, we attempt to answer four specific questions in two separate experiments. In the first experiment, we ask a) if intensity ratings for orientation items and sampled stimuli vary across the gLMS and gVAS, and b) whether the specific wording in the participant instructions influences categorical behavior on the gLMS. In the second experiment, we c) assess perceived ease of use by participants and d) compare the discrimination among samples (resolving power) between scales.

2. Methods

2.1 Participants

Study participants were recruited from the Pennsylvania State University campus and surrounding community (University Park, PA). Participants were screened to ensure they were between 18–45 years old, had not smoked in the last 30 days, had no lip or tongue piercings, no known taste defects and were not ill due to a cold or the flu. Some of the participants had prior experience in taste studies, but no participants in either experiment had used the gLMS or gVAS previously, and Experiment 1 participants were not allowed to take part in Experiment 2. Also, while we did not collect participant ages, we note that our participants are drawn from a large opt-in database maintained by our laboratory. This database consists of large number of age diverse individuals who have previously expressed an interest in routine consumer product testing in our facility (i.e., it is not a typical undergraduate psychology study pool). Procedures were exempted from Institutional Review Board review by the Penn State Office of Research Protections staff under the wholesome foods/approved food additives exemption in 45 CFR 46.101(b)(6). Participants provided informed consent and were compensated for their time.

2.2 Scaling and sample presentation

The experimenter personally greeted participants at the beginning of the test; all orientation and testing occurred in an isolated testing booth without further interaction. All scales were presented as horizontal lines on a computer screen using Compusense® *five* software version 5.2 (Compusense, Inc, Guelph ON). Prior to rating any sampled stimuli, participants were provided written instructions on how to use the scale (described below), and oriented to the scale by rating 15 imagined sensations that included a range of intensities, and sensory domains (touch, taste, tactile, thermal, auditory, visual). The orientation items were presented in a fixed order. After the orientation, participants rated the perceived intensity of a series of taste and chemesthetic solutions. All samples consisted of 10mL aliquot presented in a plastic soufflé cup at room temperature (~20°C). Sample presentation orders were randomized, and participants rinsed *ad libitum* with room temperature reverse osmosis (RO) water prior to the first sample and between each sample. For Experiment 1, participants were asked to rate the sweetness, sourness, saltiness, bitterness, umami/savory, and burning/stinging of 4 stimuli: 0.5M sucrose (Domino), 0.41mM quinine hydrochloride (SAFC), 0.56mM potassium chloride (Spectrum), and 25uM natural capsaicin (Aldrich). For experiment 2, participants rated the overall intensity of five sucrose solutions: 0.19M, 0.24M, 0.30M, 0.37M and 0.47M.

2.3 Experiment One Design and Scale Instructions

Participants (n=87) were randomized to one of three conditions in a between-subjects design (n=29 per condition), where the warm-up orientation items and sampled stimuli were identical across all three conditions. All data were collected in a single session. The differences between conditions were the absence/presence of internal semantic labels on the scale (gLMS versus gVAS) and the written instructions that explained the scale (implicit gLMS versus explicit gLMS). For both gLMS conditions, participants used a horizontal scale with ticks and semantic labels at the following points: no sensation (0) barely detectable (1.4), weak (6), moderate (17), strong (35), very strong (51), strongest imaginable sensation of any kind (100). The gVAS condition eliminated all internal labels, retaining no sensation (0) and strongest imaginable sensation of any kind (100). In the explicit gLMS condition, participants received instructions identical to those provided by Green (2002). In the implicit gLMS condition, participants received written instructions that were identical to the gVAS instructions. The exact wording is provided in Supplement 1. Evidence from Bartoshuk and colleagues (Bartoshuk, et al., 2006) indicates that the modifier *imaginary* is unnecessary in the top anchor; in spite of this, we retained ‘strongest imaginable sensation of any kind’ for all three scales to ensure greater comparability with prior reports.

2.4 Experiment Two Design and Scale Instructions

Participants were tested in a within-subjects crossover design. The scale layout and structure (gLMS and gVAS) were identical to those in Experiment 1. Condition assignment was randomized and counterbalanced so half the participants received the gLMS condition first and the other half received the gVAS condition first. The two sessions were scheduled a week apart and 58 participants completed both days of testing. In Experiment 2, the implicit gLMS instructions from Experiment 1 were used, as this allowed us to use identical wording for both the gVAS and gLMS. Perceived ease of use was also assessed (described in detail below).

2.5 Data Analysis

Data were analyzed using SAS 9.2 (Cary, NC). Regression analyses were conducted via *proc reg*. Repeated-measures analysis of variance (ANOVA) were performed via *proc mixed*, with participants as a random effect, assuming compound symmetry for the covariance structure. Adjustments for multiple comparisons were made via the Tukey-Kramer method unless otherwise noted. Significant criterion was set at alpha = 0.05. Underlying distribution of responses (kernel density estimates) were generated via *proc kde* using the default bandwidth selection options (the Sheather–Jones plug-in [SJPI] method). Categorical behavior was quantified using a method described previously (Lawless, Popper, et al., 2010). All data were used; no participants were dropped from the analyses.

3. Results

3.1 Experiment One Results

When comparing the orientation items across scales, the group means were highly correlated; this was expected as the relative ordering of items was preserved irrespective of scale type (Supplemental Figure S1). However, repeated measures mixed model ANOVA confirmed the intensity ratings of orientation items differed systematically across scale type (Figure 1). Notably, raw data on the gVAS were consistently higher than both versions of the gLMS, and this deviation was more apparent for more intense sensations. In contrast, raw data on the short and long versions of the gLMS did not differ, except for the two most extreme items (daggers in Figure 1). In a separate ANOVA (not shown) comparing just the implicit and explicit gLMS raw data across orientation items, scale instructions did not have a significant interaction (p=0.08) or main effect (p=0.81).

For the sampled stimuli, a similar pattern of higher ratings on the gVAS was observed; the primary quality ratings of sucrose, quinine, potassium chloride, and capsaicin were all higher on the gVAS than either version of the gLMS (Figure 2). In contrast, intensity ratings of the primary qualities did not differ between the two versions of the gLMS. Potassium chloride and capsaicin also had secondary taste qualities (bitter side tastes) that were greater than zero, but these ratings did not differ across scale type.

Given the results in Figure 2, we then standardized each participant’s raw data for the sampled stimuli against their rating for the brightest light they had ever seen, and reran the same analyses. After standardization, mean ratings did not differ across scale type for sucrose sweetness [F(2,84)=2.25; p = 0.11], quinine bitterness [F(2,84)=2.47; p = 0.09], capsaicin burn [F(2,84)=2.36; p = 0.10], potassium chloride bitterness [F(2,84)=2.63; p = 0.08], and capsaicin bitterness [F(2,84)=0.82; p = 0.44]. The main effect of scale was marginally significant for the saltiness of potassium chloride [F(2,84)=3.18; p = 0.047], but the mean gVAS ratings were not significantly higher than either version of the gLMS (Tukey-Kramer p’s > 0.2). In summary, when raw data are considered, the gVAS produces higher values than the gLMS when raw data are considered, but these differences disappear when data are standardized to a cross-modal reference like the brightest light ever seen.

Figure 3 shows that categorical behavior near the semantic labels for both versions of the gLMS; ratings made on the gVAS did not exhibit this pattern. To further quantify this behavior, we applied the method used previously (Lawless, Popper, et al., 2010): ratings were converted to a 200 point basis to account for half points, and a rating was considered as “categorical” if it fell ±2 units from the semantic label on the gLMS. These intervals account for 14.5% of the total space on the gLMS. With the implicit gLMS instructions, 44.4 % (193/435) of all ratings were categorical (95% Wald CI 39.8% to 49.0%); with the explicit gLMS instructions, 37.9% (165/435) of ratings were categorical (95% Wald CI 33.5% to 42.6%). Although the proportion of categorical behavior appears slightly lower with the explicit instructions (37.9% vs. 44.4%), the difference in proportions across instructions was not significant (Fisher’s exact p = 0.063), consistent with Figure 3. Additionally, Figure 3 shows a small but significant difference (Fisher’s exact p = 0.005) in the number of ratings at the very top of the implicit gLMS compared to the explicit gLMS; in either case, the proportion as a percentage of total ratings was quite low (6.7% and 2.5%, respectively). Indeed, this difference appears to be the prime driver of the nonsignificant disparity in overall categorical behavior across the two sets of instructions. If the difference at the top end is excluded from the categorical analysis, the proportion of categorical behavior becomes even more similar: 35.4% (explicit) versus 37.7% (implicit). Finally, Figure 3 also confirms that participants use more of the scale on the gVAS (ie, the ratings skew higher).

3.2 Experiment Two Results

We assessed perceived ease of use in two separate, complementary ways. First, at the end of each session, participants were instructed to “Think about the task you just completed today. On a scale of 1 to 5, with 1 being ‘strongly disagree,’ how much do you agree or disagree with the following statement?” Four separate statements were provided (Figure 4), and standard Likert descriptors were provided (i.e., strongly disagree, disagree, neither agree nor disagree, agree, and strongly agree) on the multiple-choice form. The effects of scale type, learning across days (effect of order), and their interaction were tested in mixed model repeated measures ANOVAs. The main effects of scale, after controlling for learning effects, are detailed in Figure 4; collectively they suggest the gLMS is easier to use, and that the orientation warm-up procedure is important, irrespective of scale structure. Regarding learning effects across days (order), participants were more comfortable with the sweetness rating task on the second day (mean 3.6 vs 4.0; F (1,54) = 6.87; p = .011) and they found the scale labels less confusing on the second day (mean 2.0 vs 1.8; F (1,54) = 4.17; p = .046). The other two ease of use questions did not show any learning effects (p’s < 0.5), and none of the day by scale interactions were significant (p’s <.35). At the end of the second day, we also asked participants to make a head to head forced choice regarding scale type. They were asked “Think about the scales you used [last week] and today. Which scale was easier to use?” The response options were “The scale with labels along the whole scale (weak, moderate, strong, etc)” and “The scale without any intensity labels in the middle”; a no preference option was not provided. As shown in Figure 5, 79% of participants found the gLMS easier to use (binomial p < .0001).

Experiment Two also asked participants to taste and rate the intensity of five sucrose samples ranging from 0.19 to 0.47 M. Figure 6 shows clear evidence of categorical behavior near the semantic labels for the intensity of sucrose when using the gLMS. (The kernel density estimates for 0.19M and 0.37M sucrose are similar; they were removed to reduce visual clutter on the graph). Notably, the third peak in the solid line suggests some participants rated the 0.24M sucrose solution at the midpoint between moderate and strong. A distinct shoulder is also seen near this point for the 0.3M sucrose ratings (dashed line).

When comparing the intensity ratings across the five sucrose concentrations (Figure 7), there were significant main effects of scale type and concentration, but the 2-way interaction was not significant. As occurred in Experiment 1, intensity ratings were higher when participants used the gVAS to report perceived intensity. For concentration, intensity ratings grew as concentration increased, as would be expected.

We explored the issue of resolving power (i.e. the ability to separate two close but distinct objects) in two ways. First, we compared each concentration to its nearest neighbor via a series of t-tests. As shown in Figure 7, 4 of 5 comparisons were significant for the gLMS task versus 5 of 5 when using the gVAS. Conservatively, this suggests the gVAS is no worse than the gLMS in distinguishing between samples. Second, we compared the F-ratios for the fixed effect of concentration across the two scales using the same method as Jaeger and Cardello (Jaeger & Cardello, 2009). Specifically, we performed separate one-way ANOVAs for the gVAS and gLMS data, and compared the F-ratios of the fixed effect of concentration across the two scales. To allow for a statistical comparison of these two F-values, we generated two distributions of F-values (one for each scaling method) using existing data. These two distributions were created by rerunning the same ANOVA model over and over, removing a different participant each time (ie, a leave one out approach). These two distributions were then compared via an independent sample t-test. The average F for the gLMS (48.35±0.86SD) was greater than the average gVAS F-value (40.73±0.93SD) (t_{114}=45.8; p<.0001). However, we also note that resolving power was uniformly high, regardless of scale type (the p's for all 116 leave one out models were <.0001). This suggests both scales were quite capable of separating the stimuli used here.

4. Discussion

Here, we find that the gVAS produces data similar to, but not identical with, data generated with the gLMS. Raw data obtained with the gVAS were consistently higher than ratings collected with the gLMS, and this was true for both imagined sensations during the orientation procedure and for sampled tastants and irritants. However, these differences disappeared when data were standardized to a cross-modal reference. Consistent with anecdotal reports, gLMS data were not normally distributed, as participants exhibited substantial categorical behavior, clustering their responses near the verbal labels. Moreover, providing explicit written instructions to rate between the adjectives was *not* successful in reducing this behavior. In contrast, gVAS data did not show any evidence of categorical behavior, as the gVAS lacks semantic labels to cluster around. Naïve participants in a university setting clearly preferred the gLMS over the gVAS, reporting that the gLMS was easier to use. In terms of resolving power, there was no clear advantage, as both scales allowed participants to differentiate between sucrose samples.

4.1 Wider range of scale usage on the gVAS

The correlation of group means for the gVAS and the gLMS orientation items was very high (r=0.98). This would be expected, as the rank order of items was preserved across methods. Indeed, prior work comparing the Labeled Affective Magnitude (LAM) scale to other scales report similarly large effects (r's >.94) (Lawless, Popper, et al., 2010; Schutz & Cardello, 2001). However, this does not imply the scales generate *equivalent* data, as multiple analytical approaches for both of the experiments consistently indicate ratings were higher on the gVAS (Figures 1–3, 7 and S1). Additionally, Figure S1 suggests the scale values deviated more as the intensity of the sensation increases. While it is well accepted that using a generalized scale with an extreme top anchor will compress the range of the scale that is used by participants (eg, (Cardello, et al., 2008; Ludy & Mattes, 2011)), we should emphasis that unlike recent work by Ludy and Mattes, both of the scales here used an identical top anchor. Thus, it appears that the internal labels themselves (eg, very strong) also cause some degree of compression. That is, the presence of a tick at 51 with a label of 'very strong' may encourage participants to make the majority of their ratings in the bottom half of the scale. Additionally, while gVAS values were consistently higher than gLMS values, the gVAS generates ratio level data similar to the gLMS (e.g., 300mM sucrose was ~1.9 times more intense than 190mM sucrose, regardless of which scale is used. Likewise, the 470mM sucrose was ~1.95 times more intense than the 240mM sucrose on both scales.)

There are two schools of thought regarding compression and scale performance. One school holds that compression is inherently undesirable, as it is theoretically preferable for participants to use as much of the scale as possible to maximize the ability to distinguish between samples (Cardello, et al., 2008; Lawless, Cardello, et al., 2010). In practice however, compression is often accompanied by smaller variance, which can preserve the ability to find differences between products (Lawless, Popper, et al., 2010). Alternatively, another line of thinking suggests compression resulting from a generalized top anchor is not worrisome (or may even be desirable) if it reduces ceiling effects. Indeed, generalized top anchors seem to improve the ability to make valid comparisons across individuals (Snyder, et al., 2004) who differ in physiology or genetics (eg Hayes, Bartoshuk, Kidd, & Duffy, 2008). Likewise, in affective testing, generalized top anchors seem to reduce ceiling effects for highly liked products (Lim & Fujimaru, 2010; Schutz & Cardello, 2001). Given this, we do not recommend one scale over the other universally, suggesting instead that the choice of scale is contingent on the goals and constraints of the specific experiment. However, we would caution that researchers used to gLMS data should be extremely careful about unintentionally inferring a certain response magnitude when working with gVAS data (eg, a 35 on the gVAS is not ‘strong’).

4.2 Categorical Behavior on the gLMS

Clustering of responses near the semantic labels (“categorical behavior”) is commonly known among those who use the gLMS – here, we provide formal evidence of this phenomenon. Lawless and colleagues recently quantified categorical behavior on another generalized scale, the Labeled Affective Magnitude (LAM) scale (Cardello, et al., 2008; Lawless, Popper, et al., 2010; Lawless, Sinopoli, et al., 2010). Subsequently, Lim and Fujimaru (2010) reported similar results with the LHS. An unstructured generalized hedonic scale, the SLAM (simplified LAM), was recently described (Lawless, Cardello, et al., 2010); one assumes it lacks the categorical behavior found on the LAM and LHS. Here, the distinct peaks in Figures 3 and 6 clearly indicate categorical behavior is not limited to affective scaling, as it also occurs when participants make intensity ratings using the gLMS. Although the gLMS is often described as generating log-normal data, present data are obviously multimodal. Nor is categorical behavior an artifact based on the specific remembered sensations used in the orientation procedure, as it occurs in our data for sampled sucrose. We have checked other datasets and found the same pattern for salty stimuli as well (not shown). Notably, present data also indicate that providing explicit written instructions to participants to use the space in between the semantic labels did not significantly reduce this behavior. It is conceivable that explicit *verbal* instructions from the experimenter may be more effective at reducing this behavior than written instructions. Analysis of other datasets where participants received verbal instructions suggests categorical behavior still occurs (not shown), but we cannot make direct comparisons to present data due to other differences in testing conditions. This should be formally tested. Additionally, we note that the need to perform detailed one on one training would substantially limit the utility of generalized scales in high throughput testing situations.

How categorical behavior influence data analysis is unknown. The lack of this behavior on the gVAS may be wholly irrelevant if such clustering does not compromise the statistical models that are subsequently created with these data. Additionally, it should be obvious that removing the internal labels may be undesirable due to the loss of semantic information that may be otherwise useful. If the goal is merely to compare product formulations, numeric values from the gVAS may be sufficient (eg, a sweetness of 20 and 30 for products A and B will tell you B is 50% sweeter). However, if an experimenter wants to translate a mean scale value into a term with more communication value (eg, the burn was moderate, and not strong), clearly, the gLMS should be used.

4.3 Ease of Use

Contrary to our intuition, the participants in our study clearly preferred the gLMS to the gVAS. We also asked ease of use questions that related to both the task difficulty and the scale itself; these were generally consistent with the head to head preference. Our participants uniformly felt the scale orientation procedure helped them understand how to use the scale, regardless of the scale structure. We are tempted to speculate that many of the purported advantages of generalized scales may be due to the cross modal orientation protocol and not some special properties inherent to the wording of the top anchor. If this is confirmed, this would suggest it is critical for experimenters to include a warmup orientation procedure when they wish to collect data with a generalized scale, be it the gVAS or gLMS.

In practice, we previously observed that the adjectives on the gLMS confuse some participants, since the same semantic label may denote different absolute intensities depending on the specific sensory domain being discussed. For example, a woman who has just experienced childbirth may describe her pain as being ‘very strong.’ She may also describe the odor of the roses in a bouquet presented to her to celebrate the birth as ‘very strong’. Colloquially, we understand that she does not intend to equate their intensity (Duffy, Hayes, Bartoshuk, & Snyder, 2009), so we had assumed this elasticity made the scale instructions confusing for some participants. In addition to S.S Stevens’ famous quotation regarding mice and elephants (1958), the problem of domain specificity of semantic labels, and its variability across people, has been discussed within the applied sensory literature (see Munoz & Civille, 1998). However, the proposed solution there (universal scaling) relies on references that remain largely within a consumer products/ chemosensory context, so Bartoshuk’s criticisms would still apply, if the goal was to make comparisons across people instead of across products. To our surprise, it appears that a lack of an internal structure and references was more troubling to participants than adjective elasticity, as they indicated the rating task was harder in the absence of internal ticks and labels on the scale. Whether this would hold true in other groups is unknown, as the present ease of use data may not extend to other populations. Here, we tested a literate, numerate cohort of adults in university setting. Whether these data extend to children (eg, Timpson, et al., 2007), or other special populations like the elderly, substance abusers, or community based samples is unknown. Notably, the first usage of the gVAS in the literature was in 10 year old children (Timpson, et al., 2007); presumably, ease of use was the motivation, although this was not directly assessed in that study.

4.4 Resolving Power

The issue of relative resolving power across the two scales is difficult to answer; in attempting to address this question in multiple ways, we observed conflicting results. Based on the size of the F-ratios for the sample (concentration) effect, it would appear that discrimination between samples was slightly superior with the gLMS. Conversely, based on the ability of each scale to maximize the number of significant differences between neighboring concentrations, the gVAS was as least as good, and possibly better than the gLMS. Curiously, the one case in which the gVAS found a difference that the gLMS did not was between the two highest concentrations. At risk of over interpreting a single datum, it is temping to note that this is exactly what one might expect from compression, if the internal anchor ‘very strong’ acts as a false ceiling that pushes ratings of sampled stimuli down into the bottom half of the scale.

The sucrose concentrations used here were not intended to be easy to discriminate. While some earlier reports had suggested a larger value, newer work indicates the Weber Ratio for sucrose is ~0.13 (Gilmore & Murphy, 1989; Mcbride, 1983). To accommodate the nature of the task (direct scaling vs. a forced choice paradigm), we inflated this by value by 50%, obtaining an adjusted ratio of 0.2. Thus, from a starting point of 0.3M sucrose, the other concentrations represent steps up or down of 20%. Based on present data, it appears our inflation may have been overly generous, as the F-ratios were quite large, regardless of scale type. More work, with more diverse product sets and sample spacing, is needed to determine if one scale is in fact superior to the other in discrimination ability. For now, we would simply suggest that the gVAS is no worse than the gLMS is distinguishing between samples when there is moderate separation between samples.

4.5 Summary and Conclusions

Here, the orientation procedure encouraged participants to use the full range of the unstructured gVAS scale, with both order and spacing of items similar to the traditional gLMS. Moreover, it did so without clustering artifacts (ie, categorical behavior). That said, the gVAS gives up the semantic information provided by the gLMS. As we have noted previously (Duffy, et al., 2009), researchers may find adjective labels especially useful in communicating intensities as words rather than numbers (eg, ‘the burn was strong’ vs. ‘the burn was 37’). The ability to do so is lost with the gVAS. Also, we note that use of a generalized scale may be inappropriate for product testing if scale compression obfuscates differences between products (eg, Lawless, Cardello, et al., 2010), although other reports suggests generalized scales may increase discriminability by reducing ceiling effects (eg, Schutz & Cardello, 2001).

In summary, present data indicate chemosensory researchers should be careful in how they interpret raw data collected with the gVAS (eg, Pickering, et al., 2010; Timpson, et al., 2007). Both the gVAS and gLMS generate data that discriminate between stimuli, and preserve order among stimuli, but the raw numeric values are not one to one comparable. Present data also highlight the importance of the orientation procedure, regardless of scale structure. Our adult cohort found the gLMS easier to use; whether this is the case in children or other special populations is unknown. Present work suggests that the choice of one scale over the other depends on whether the experimenter believes the lack of categorical behavior offsets the loss of semantic information in his or her project.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by National Institute of Health grant (R03DC010904) to JEH. The authors thank the study participants for their time and participation, and Dr. David DePasquale.

References

Aitken RC. Measurement of feelings using visual analogue scales. Proc R Soc Med. 1969; 62(10): 989–993. [PubMed: 4899510]

Bartoshuk LM. Comparing sensory experiences across individuals: recent psychophysical advances illuminate genetic variation in taste perception. Chem Senses. 2000; 25(4):447–460. [PubMed: 10944509]

Bartoshuk LM, Duffy VB, Fast K, Green BG, Prutkin J, Snyder DJ. Labeled scales (eg, category, Likert, VAS) and invalid across-group comparisons: what we have learned from genetic variation in taste. Food Qual Pref. 2003; 14(2):125–138.

Bartoshuk LM, Duffy VB, Hayes JE, Moskowitz HD, Snyder DJ. Psychophysics of sweet and fat perception in obesity: problems, solutions and new perspectives. Philosophical Transactions of the Royal Society B: Biological Sciences. 2006; 361(1471):1137–1148.

Cardello A, Lawless HT, Schutz HG. Effects of extreme anchors and interior label spacing on labeled affective magnitude scales. Food Quality and Preference. 2008; 19:473–480.

Dionne RA, Bartoshuk L, Mogil J, Witter J. Individual responder analyses for pain: does one pain scale fit all? Trends in Pharmacological Sciences. 2005; 26(3):125–130. [PubMed: 15749157]

Duffy, VB.; Hayes, JE.; Bartoshuk, LM.; Snyder, DJ. Taste: Vertebrate Psychophysics. In: Squire, LR., editor. Encyclopedia of Neuroscience. Oxford: Academic Press; 2009. p. 881-886.

Duffy VB, Peterson J, Bartoshuk LM. Associations between taste genetics, oral sensations and alcohol intake. Physiol Behav. 2004; 82(2–3):435–445. [PubMed: 15276808]

Eggleston K, White TL, Sheehe PR. Adding cocoa to sucrose: the effect on cold pain tolerance. Chem Senses. 2010; 35(4):269–277. [PubMed: 20197300]

Gilmore MM, Murphy C. Aging Is Associated with Increased Weber Ratios for Caffeine, but Not for Sucrose. Perception & Psychophysics. 1989; 46(6):555–559. [PubMed: 2587184]

Green, BG. Psychophysical Measurement of Oral Chemesthesis. In: Simon, SA.; Nicolelis, MAL., editors. Methods in chemosensory research. Boca Raton, FL: CRC Press; 2002. p. 527

Green BG, Dalton P, Cowart B, Shaffer G, Rankin K, Higgins J. Evaluating the 'Labeled Magnitude Scale' for measuring sensations of taste and smell. Chem Senses. 1996; 21(3):323–334. [PubMed: 8670711]

Green BG, Hayes JE. Capsaicin as a probe of the relationship between bitter taste and chemesthesis. Physiology & Behavior. 2003; 79(4–5):811–821. [PubMed: 12954427]

Green BG, Shaffer GS, Gilmore MM. Derivation and evaluation of a semantic scale of oral sensation magnitude with apparent ratio properties. Chem Senses. 1993; 18(6):683–702.

Hayes JE, Bartoshuk LM, Kidd JK, Duffy VB. Supertasting and PROP Bitterness Depends on More Than the TAS2R38 Gene. Chem Senses. 2008; 33(3):255–265. Epub 221 January 2008. doi: 2010.1093/chemse/bjm2084. [PubMed: 18209019]

Hayes JE, Duffy VB. Revisiting sugar-fat mixtures: sweetness and creaminess vary with phenotypic markers of oral sensation. Chem Senses. 2007; 32(3):225–236. First published January 224, 2007. doi:2010.1093/chemse/bjl2050. [PubMed: 17204520]

Hayes, JE.; McGeary, JE.; Grenga, A.; Swift, RM. Do TAS1R3 promoter region SNP rs35744813 A allele carriers show a reduced response to concentrated sucrose? [ABSTRACT]. AChemS XXXII; St. Pete's Beach, Florida. 2010.

Hayes MH, Patterson DG. Experimental Development of the Graphic Rating System. Psychol Bull. 1921; 18(2):98–99.

Jaeger SR, Cardello AV. Direct and indirect hedonic scaling methods: A comparison of the labeled affective magnitude (LAM) scale and best-worst scaling. Food Quality and Preference. 2009; 20(3):249–258.

Keast RS, Roper J. A complex relationship among chemical concentration, detection threshold, and suprathreshold intensity of bitter compounds. Chem Senses. 2007; 32(3):245–253. [PubMed: 17220518]

Lawless HT, Cardello AV, Chapman KW, Lesher LL, Given Z, Schutz HG. A Comparison of the Effectiveness of Hedonic Scales and End-Anchor Compression Effects. Journal of Sensory Studies. 2010; 25:18–34.

Lawless HT, Malone GJ. Comparison of Rating Scales: Sensitivity, Replicates and Relative Measurement. Journal of Sensory Studies. 1986a; 1(2):155–174.

Lawless HT, Malone GJ. The Disriminative Efficiency of Common Scaling Methods. Journal of Sensory Studies. 1986b; 1(1):85–98.

Lawless HT, Popper R, Kroll BJ. A comparison of the labeled magnitude (LAM) scale, an 11-point category scale and the traditional 9-point hedonic scale. Food Quality and Preference. 2010; 21(1): 4–12.

Lawless HT, Sinopoli D, Chapman KW. A Comparison of the Labeled Affective Magnitude Scale and the 9-Point Hedonic Scale and Examination of Categorical Behavior. Journal of Sensory Studies. 2010; 25:54–66.

Lim J, Fujimaru T. Evaluation of the Labeled Hedonic Scale under different experimental conditions. Food Quality and Preference. 2010; 21(5):521–530.

Lim J, Wood A, Green BG. Derivation and evaluation of a labeled hedonic scale. Chem Senses. 2009; 34(9):739–751. [PubMed: 19833660]

Ludy MJ, Mattes R. Noxious Stimuli Sensitivity in Regular Spicy Food Users and Non-Users: Comparison of Visual Analog and General Labeled Magnitude Scaling. Chemosensory Perception. 2011; 4(4):123–133.

Marks LE, Stevens JC, Bartoshuk LM, Gent JF, Rifkin B, Stone VK. Magnitude-matching: the measurement of taste and smell. Chemical Senses. 1988; 13(1):63–87.

Mcbride RL. A Jnd-Scale Category-Scale Convergence in Taste. Perception & Psychophysics. 1983; 34(1):77–83. [PubMed: 6634362]

Moskowitz H. Magnitude estimation: Notes on what, how, when, and why to use it. Journal of Food Quality. 1977; 1(3):195–227.

Munoz AM, Civille GV. Universal, product and attribute specific scaling and the development of common lexicons in descriptive analysis. J Sens Stud. 1998; 13(1):57–75.

Pickering GJ, Moyes A, Bajec MR, Decourville N. Thermal taster status associates with oral sensations elicited by wine. Australian Journal of Grape and Wine Research. 2010; 16(2):361–367.

Schutz HG, Cardello AV. A labeled affective magnitude (LAM) scale for assessing food liking/ disliking. Journal of Sensory Studies. 2001; 16(2):117–159.

Snyder DJ, Fast K, Bartoshuk LM. Valid Comparisons of Suprathreshold Sensations. Journal of Consciousness Studies. 2004; 11(7–8):96–112.

Snyder DJ, Prescott J, Bartoshuk LM. Modern psychophysics and the assessment of human oral sensation. Adv Otorhinolaryngol. 2006; 63:221–241. [PubMed: 16733341]

Stevens SS. Adaptation-Level Vs the Relativity of Judgment. American Journal of Psychology. 1958; 71(4):633–646. [PubMed: 13627273]

Timpson NJ, Heron J, Day IN, Ring SM, Bartoshuk LM, Horwood J, Emmett P, Davey-Smith G. Refining associations between TAS2R38 diplotypes and the 6-n-propylthiouracil (PROP) taste test: findings from the Avon Longitudinal Study of Parents and Children. BMC Genet. 2007; 8(1): 51. [PubMed: 17662150]

Zealley AK, Aitken RC. Measurement of mood. Proc R Soc Med. 1969; 62(10):993–996. [PubMed: 5346176]

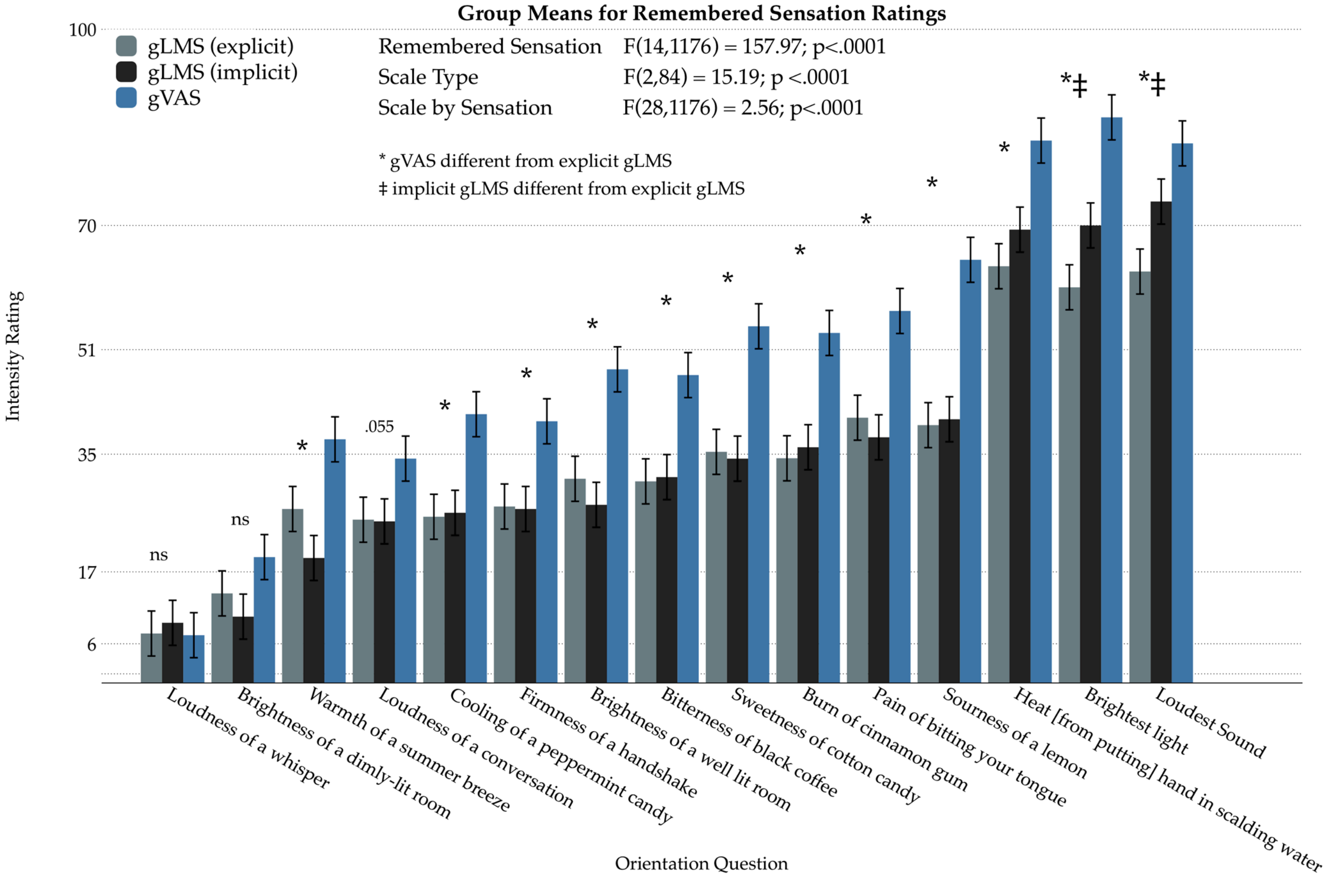

Figure 1.

Differences in means for remembered and imagined sensations collected with the gLMS (light and dark grey bars) and the gVAS (blue bars) for Experiment 1. Two sets of gLMS instructions were tested; the explicit instructions emphasized that participants should make ratings between the verbal intensity labels while the implicit instructions did not. The 15 orientation items encouraged the participants to use most of the scale range. Ratings for the gVAS were consistently higher than either version of the gLMS. Stars and daggers indicate significant Tukey adjusted posthoc comparisons at p<0.05.

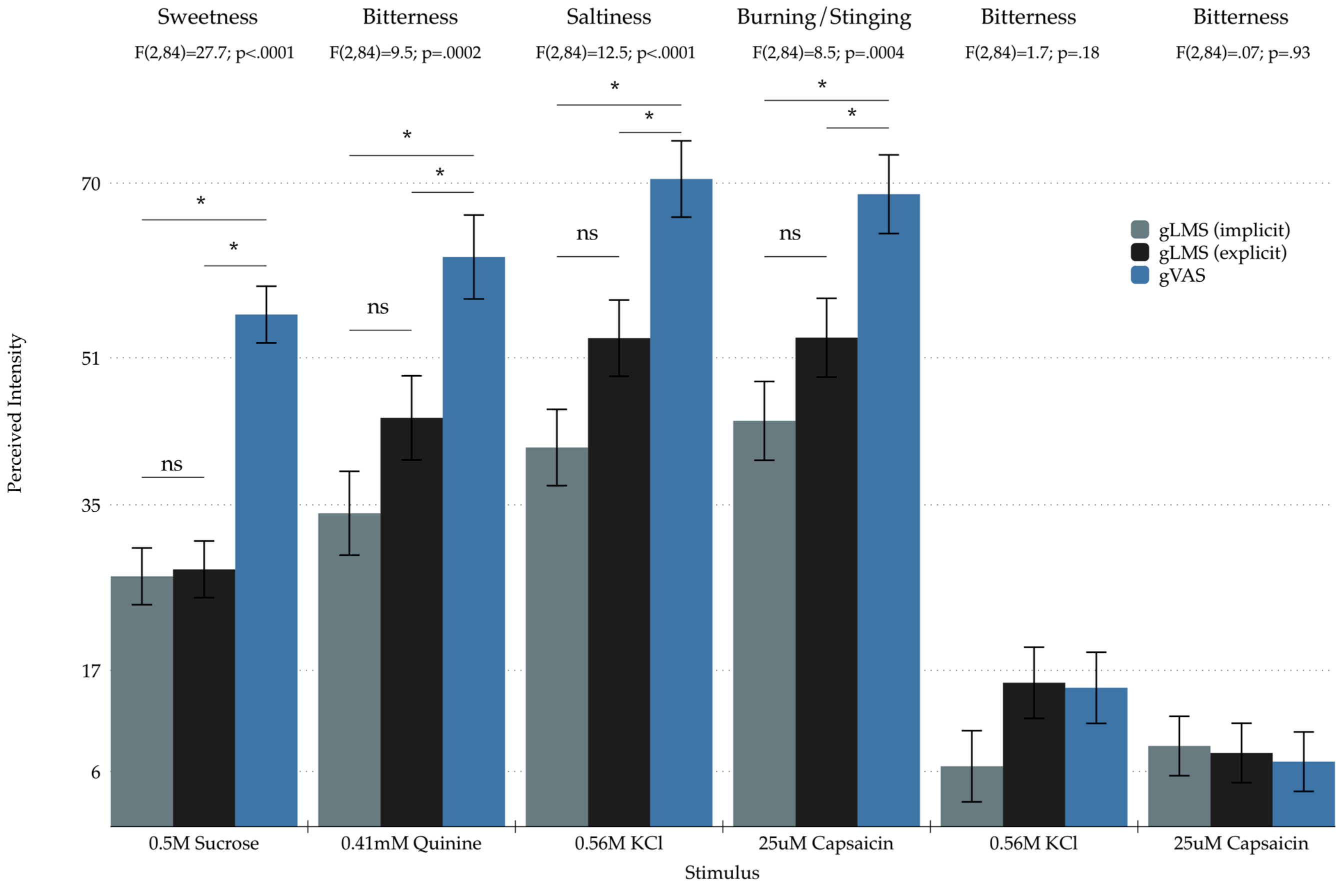

Figure 2.

Differences in means for sampled stimuli across the three scaling methods for Experiment 1. Ratings for the gVAS were consistently higher than either version of the gLMS for the primary quality of each stimulus. Potassium chloride and capsaicin exhibited significant non-zero side tastes as expected, but the magnitude of these ratings did not differ across scale type. Bar colors are the same as in Figure 1 and stars indicate significant Tukey adjusted posthoc comparisons at p<0.05.

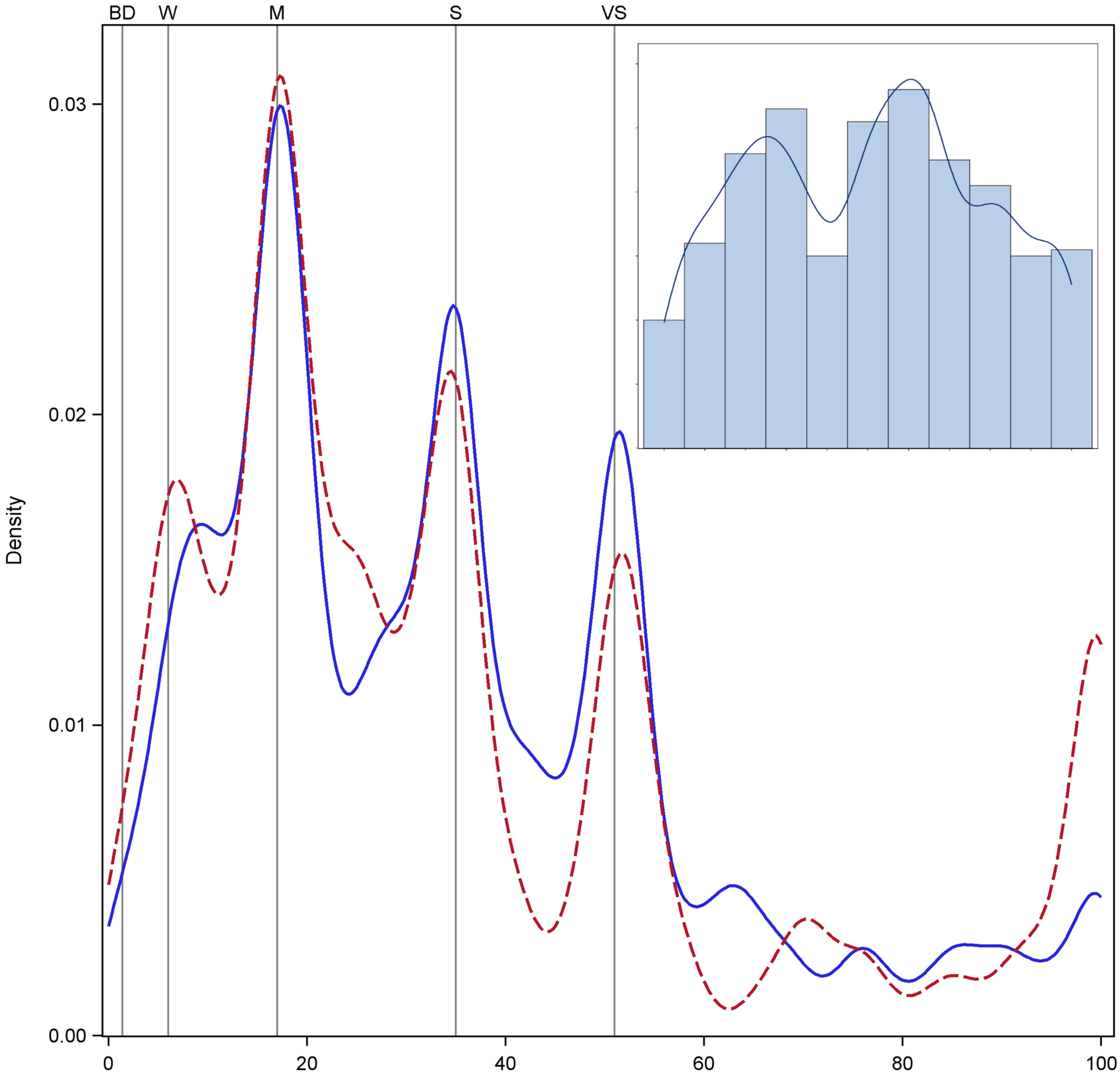

Figure 3.

Kernel density estimates of all ratings made during the orientation procedure in Experiment 1. Categorical behavior near the verbal labels is clearly seen for both versions of gLMS; this behavior was not observed in the gVAS (inset figure top right). Providing instructions that emphasized participants should rate between the verbal intensity labels (solid blue line) did not reduce this behavior compared to instructions that were otherwise identical (dashed red line).

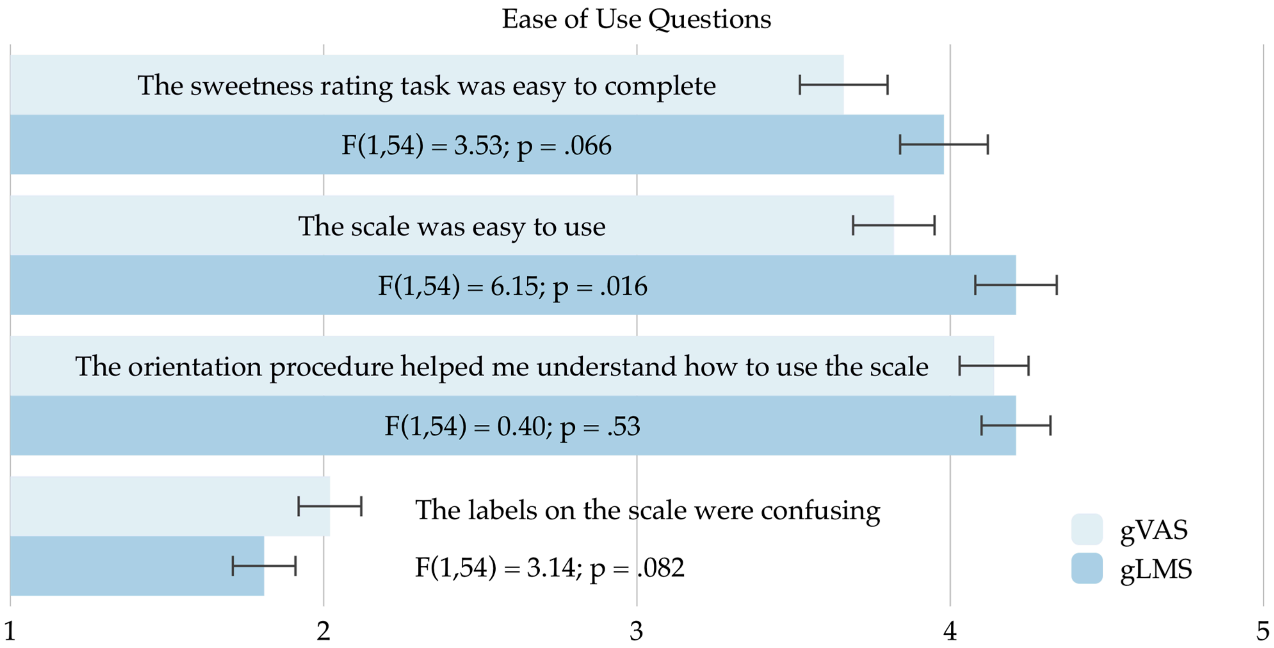

Figure 4.

Ease of use estimates from Experiment 2. The numbers on the x-axis correspond to standard Likert items, where 1 is 'strongly disagree', 2 is 'disagree', 3 is 'neither agree nor disagree', 4 is 'agree', and 5 is 'strongly agree'. Irrespective of scale type, the participants found the orientation procedure helpful. The other three questions show the gLMS was preferred by participants in a university setting. The individual F-values indicate the main effect of scale type after controlling for learning effects across sessions.

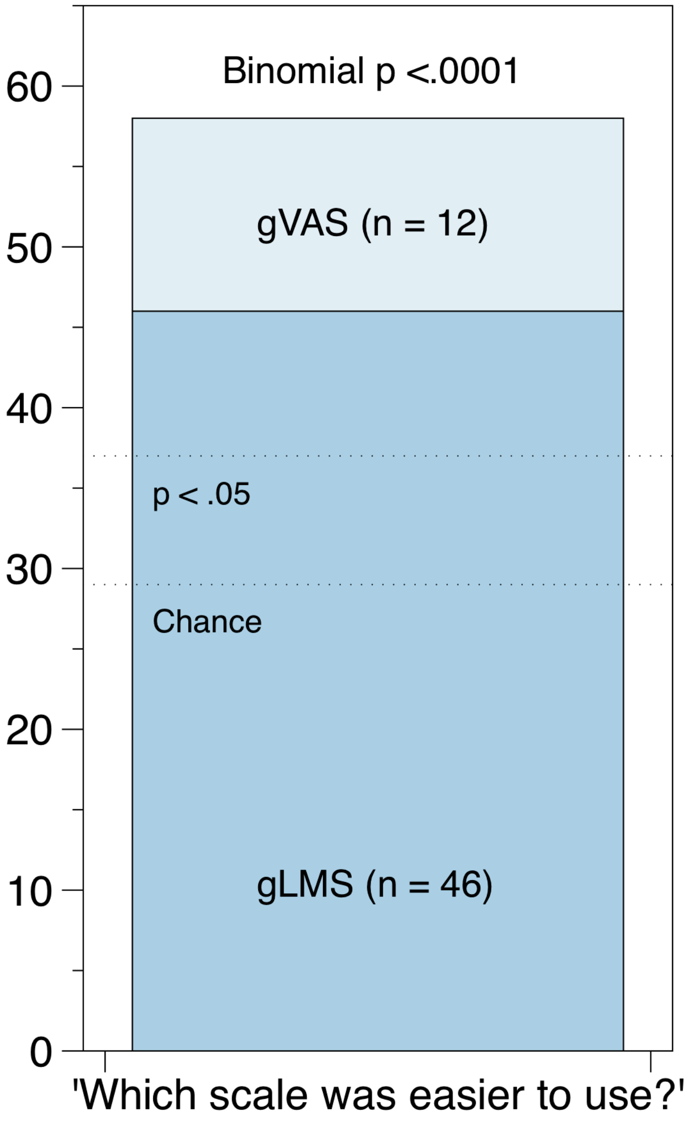

Figure 5.
In a head to head comparison at the end of the second session in Experiment 2, 79% of participants indicated the gLMS was easier to use than the gVAS.

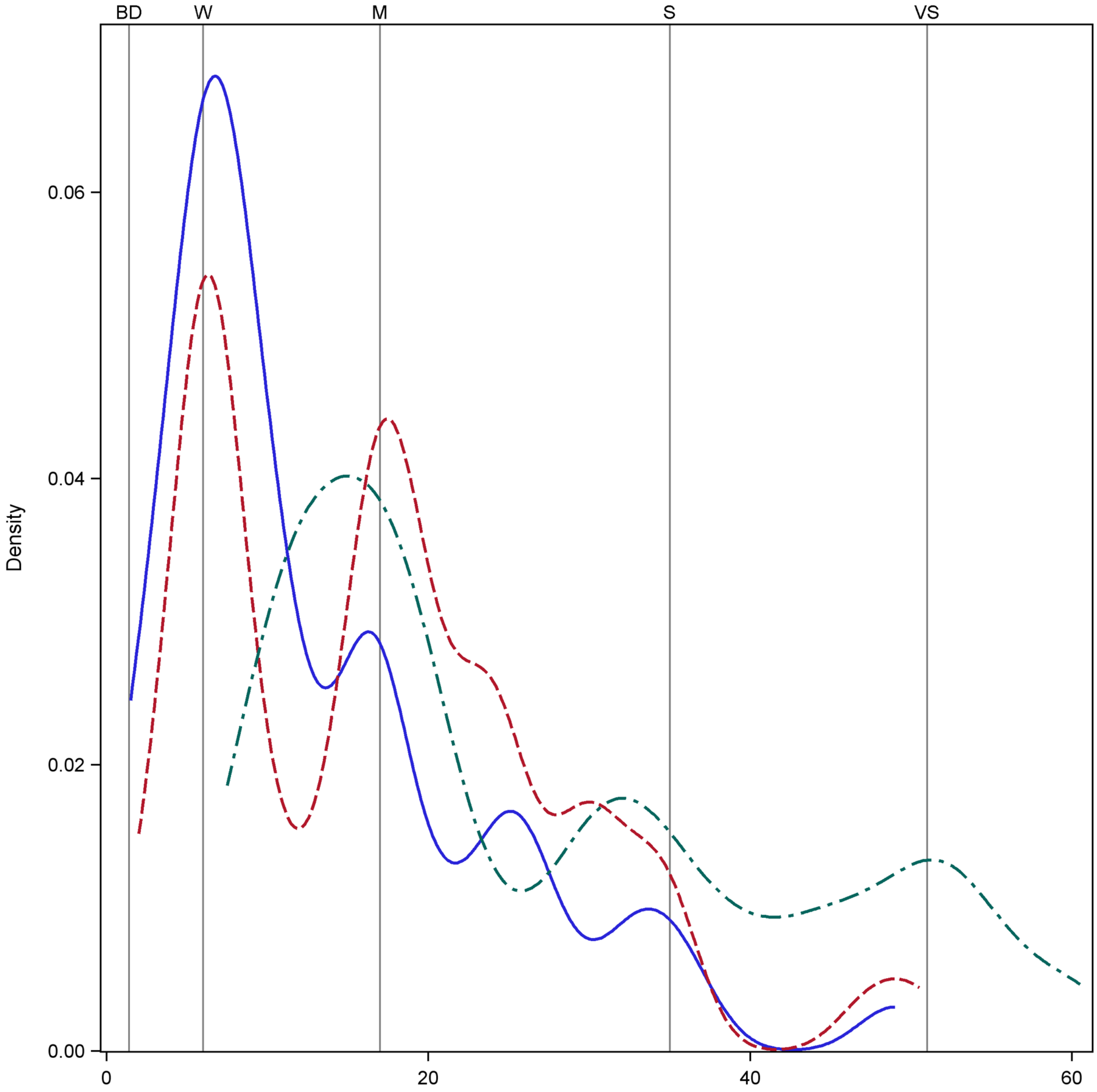

Figure 6.

Kernel density estimates for sucrose samples in Experiment 2. Categorical behavior near the verbal labels is clearly seen for the three sucrose concentrations shown (solid blue = 0.24 M; red dashed = 0.30M; green dash dot = 0.47M); the other two concentrations are similar and were removed to reduce visual clutter.

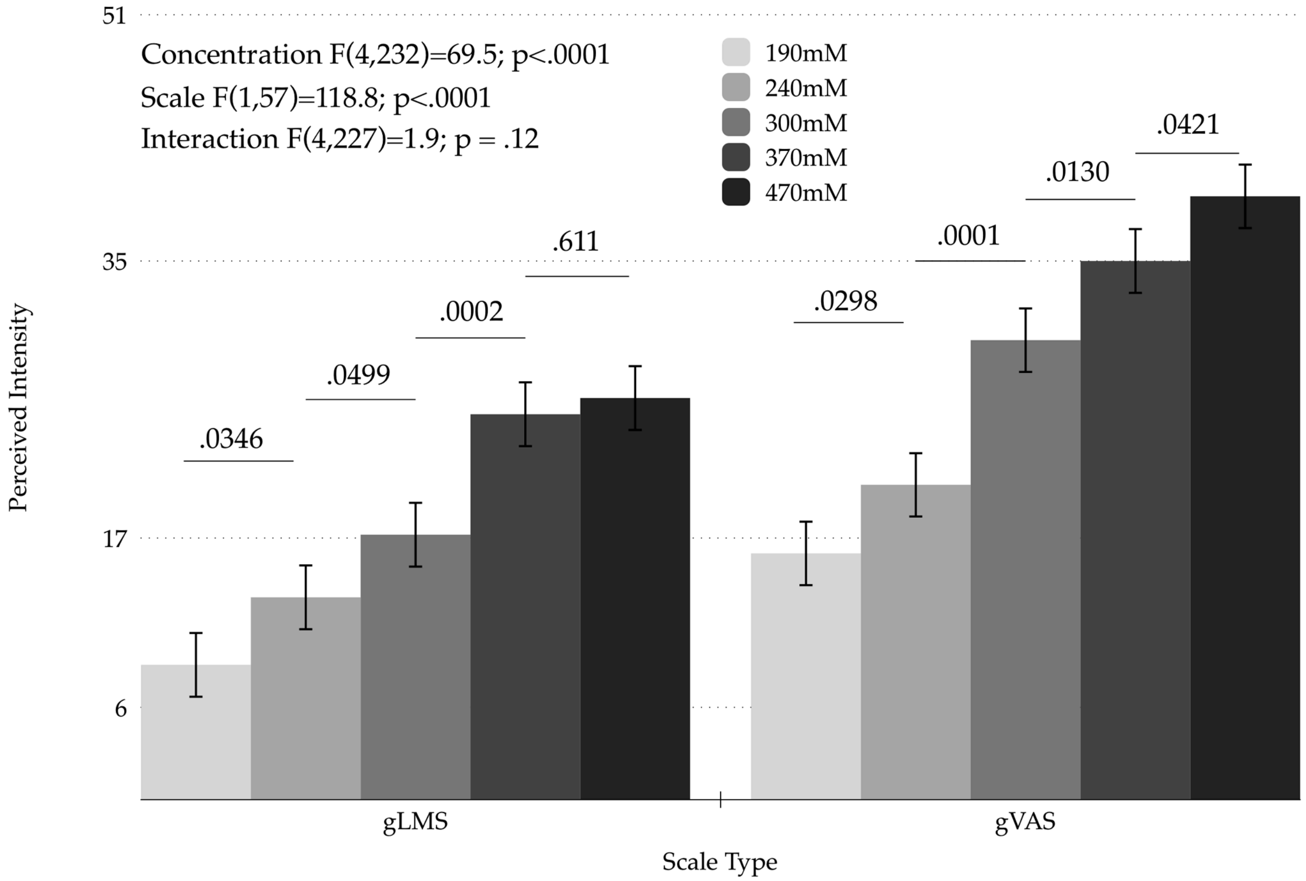

Figure 7.

Differences in means for sampled sucrose across concentration for both scaling methods in Experiment 2. As expected, ratings increased with concentration, and ratings were higher when collected with the gVAS. Significance between neighboring concentrations within a scaling method were determined with unadjusted t-tests.