2019

Data approximation strategies between generalized line scales and the influence of labels and spacing

Jonathan C. Kershaw

Cordelia Running

Data approximation strategies between generalized line scales and the influence of labels and spacing
JC Kershaw, CA Running
Journal of Sensory Studies, e12507

https://doi.org/10.1111/joss.12507

Data approximation strategies between generalized line scales and the influence of labels and spacing

Running title: Data approximation strategies

Jonathan C. Kershaw[1,2,3]

Cordelia A. Running[1,2]*

[1]Department of Nutrition Science, 700 W State St, Purdue University, West Lafayette IN USA

[2]Department of Food Science, 700 W State St, Purdue University, West Lafayette IN USA

[3]Department of Public and Allied Health, 136 Health and Human Services, Bowling Green State University, Bowling Green OH USA

*Corresponding author: crunning@purdue.edu

Abstract

Comparing sensory data gathered using different line scales is challenging. We tested whether adding internal labels to a generalized visual analog scale (gVAS) would improve comparability to a typical generalized labeled magnitude scale (gLMS). Untrained participants evaluated cheeses using one of four randomly assigned scales. Normalization to a cross-modal standard and/or two gLMS transformations were applied to the data. Response means and distributions were lower for the gLMS than the gVAS, but no difference in resolving power was detected. The presence of words, independent of internal lines, induced categorical behavior (marking close to labels). Closer low-end label spacing for gLMS increased influenced participants to mark near higher intensity labels when they were evaluating low-intensity samples. Although normalization reduced differences between scales, neither transformation nor normalization was supported as appropriate gLMS/gVAS approximation strategies. This study supports previous observations that neither scale offers a systematic advantage and that participant usage differences limit direct scale comparisons.

Practical Applications

Practitioners should exercise caution when comparing between gVAS and gLMS data, as neither normalization nor transformations make these equivalent. The value of qualitative information from internal labels (gLMS) and the expected intensity of samples should be considered when choosing a scale (gVAS may be better for lower intensity samples, gLMS for high intensity). While all scales in this study provide valid information regarding sample intensity, the impact of scale on statistical assumptions, most notably the non-normal distribution of residuals from gLMS data, should also be considered and corrected when necessary.

Key words: Scale selection, categorical behavior, transformation, normalization, gLMS, gVAS

1. Introduction

Continuous visual analog scales (VAS) are commonly used to quantify and compare sensory experiences. The VAS is a continuous line anchored at either end by a minimum and maximum, and is considered a practical and reliable tool for measuring and comparing human experiences across different populations (Price, McGrath, Rafii, & Buckingham, 1983; Zealley & Aitken, 1969). Furthermore, the continuous nature of the scale allows for more sensitive statistical tests compared with category scales. In some uses of the scale, internal semantic labels are added to provide qualitative information about sensation intensity. However, VAS are not without limitations (Bartoshuk et al., 2003). A desirable line scale would provide semantic information; produce normally distributed, ratio-level data; and have high resolving power (i.e., the ability to detect differences between distinct samples), among other desirable attributes.

In sensory research, the labeled magnitude scale (LMS) was developed to produce data with ratio-level properties and qualitative meaning, which previously required the use of either magnitude estimation or category scales, respectively (Green, Shaffer, & Gilmore, 1993). On the LMS, internal labels are spaced quasi-logarithmically based on the quantitative derivation of semantic descriptors. Building on findings of the LMS, other researchers suggested a top anchor of "the strongest imaginable sensation of any kind," to compare results across populations with systematic differences in experiences and physiology, such as genetic differences in 6-n-propylthiouracil sensitivity (Bartoshuk et al., 2003, 2004). This scale was designated as a generalized labeled magnitude scale (gLMS). Further refinement of the scale revealed that the descriptor "imaginable" does not improve across group comparisons (Bartoshuk, Fast, & Snyder, 2005). Others have applied a generalized, cross-modal top anchor to a VAS with no internal labels and designated it as the general visual analog scale (gVAS) (Bartoshuk et al., 2005). Like the gLMS, the gVAS can also produce ratio level data (Hayes, Allen, & Bennett, 2013). Of note, generalized scales improve across group comparisons but do not necessarily offer an advantage for within subject analyses (Kalva et al., 2014). Both of these generalized scales, as well as a variety of derivatives from them, are commonly used in sensory evaluation.

The presence of internal labels, the key difference between the gLMS and the gVAS, provides both a benefit and a limitation: although labels provide meaningful qualitative information, they also influence participants to rate closely to the markings (i.e., categorical behavior), thus producing clustered rather than continuous data (Hayes et al., 2013). As only one study has directly compared the gLMS and the gVAS (Hayes et al., 2013), further research is helpful for understanding how these scales compare in a variety of settings and sample sets, as well as for selecting among these scales and other potential derivatives of these scales.

Although the gLMS and gVAS use identical end anchors, systematic differences in the way participants use the scales limit inter-scale comparisons (Hayes et al., 2013). A method to approximate sample means generated by the two scales would facilitate cross-study comparisons and aid appropriate scale selection. To compare across scales, systematic transformations have been applied in a number of contexts for both intensity and hedonic ratings (Green et al., 1993; Lim, Wood, & Green, 2009; Schutz & Cardello, 2001). Normalizing individual ratings to a cross-modal standard (e.g. the brightness of the sun, the intensity of a tone, or the heaviness of jars of sand) has also been used to compare scales (Duffy, Peterson, & Bartoshuk, 2004; Hayes et al., 2013; Webb, Bolhuis, Cicerale, Hayes, & Keast, 2015). In addition, a greater understanding of how scale elements (i.e., anchors, label presence and spacing, etc.) influence participant behavior could aid scale selection and optimization.

The purpose of this study is to explore the comparability of a gLMS and gVAS, following normalization and/or transformations, and to assess how label presence and spacing influence participant responses.

2. Methods

2.1 Study participants and procedures

Healthy participants were recruited at the Indiana State Fair (n=195, 130 women, 65 men, 0 other, ages 8-80, mean age 38.2) to evaluate three cheese samples (blue cheese crumbles, white cheese cubes, and shaved parmesan cheese, donated by the American Dairy Association of Indiana). All protocols were approved as exempt by Purdue University's Institutional Review Board for Human Subjects Research under category 6, testing of foods and food ingredients. Inclusion criteria included absence of food allergies and willingness to eat and evaluate cheeses; acceptance of these criteria was done electronically and participation constituted consent. Participants first answered six cross-modal "warm-up" questions about remembered or imagined sensations to familiarize participants with the scale and to verify that they understood its usage (Table 1, a reduced form of the warm-up from Hayes et al., 2013; responses from participants that incorrectly used the scales were excluded data analyses, as described below). Following the warm-up questions, participants evaluated the three cheese samples in a counter-balanced order. A five second wait time was enforced between each sample. Participants were randomly assigned one of four scales (described in detail below) and asked to rate the odor intensity of the sample (participants were allowed to taste the samples as well, but not all participants chose to do so; thus, we will only report on the odor data). In addition to a gLMS and a gVAS, two other scales were created with the explicit purpose of investigating scale elements (Figure 1). The gVAS labeled with equally-spaced gLMS labels is designated as the “generalized labeled VAS” (glVAS), and the same scale without the internal lines is designated as the “generalized words-only VAS” (gwVAS). The purpose of adding these scales to the analysis was to see if spacing out the internal lines would improve resolving power for lower intensity ratings (glVAS), and if removing the actual tick marks from the main line would reduce clustering of the ratings (gwVAS); these scales have not been validated for general use.

Table 1. Instructions and questions presented to participants to familiarize them with generalized scales.

First we'd like to familiarize you with our rating system. You will rate the strength of several imagined or remembered sensations. Please rate the sensations to the best of your ability. We use this information to verify that you are reading the directions! If you have not experienced a particular sensation, rate how intense you imagine it is. If you don't pay attention here, we cannot use ANY of your data from the rest of the test. This makes us very sad.
The brightness of this room
The brightness of the sun on a clear day
The loudness of a shout
The loudness of a whisper
The bitterness of black coffee
The sweetness of pure sugar

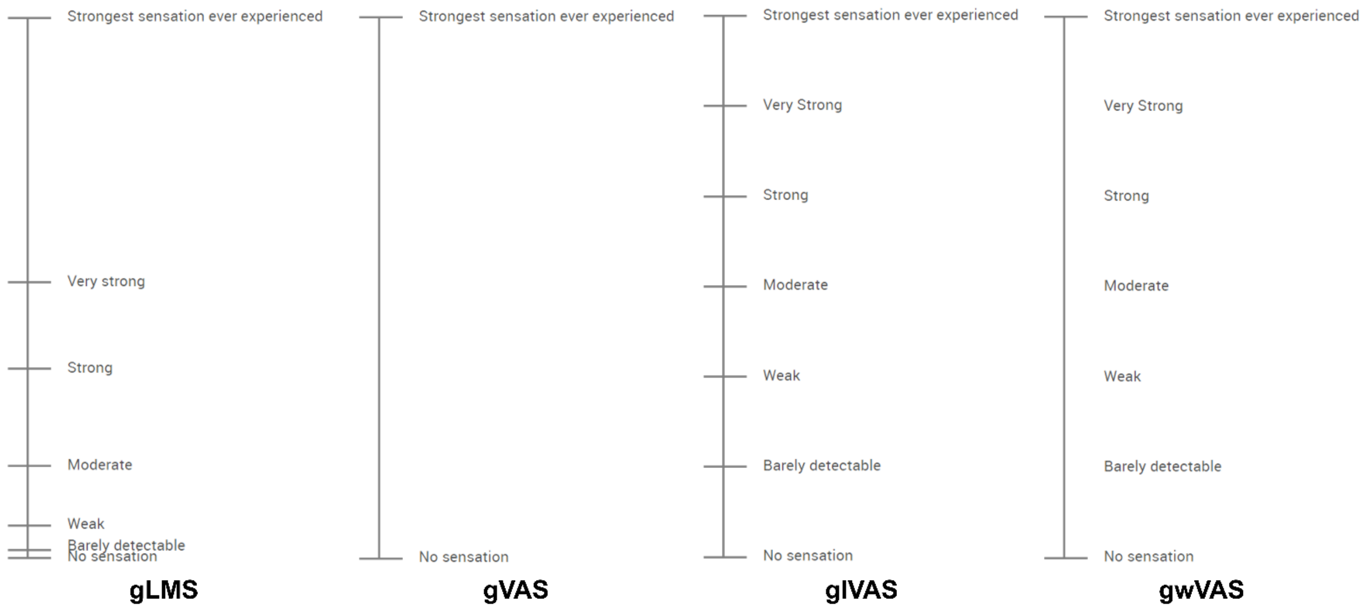


Figure 1 Scales presented to study participants. gLMS: generalized labeled magnitude scale; gVAS: generalized visual analg scale: glVAS: generalized labeled visual analog scale; gwVAS: generalized words-only visual analog scale.

2.2 Transformations and normalization

To compare responses generated by the gLMS and the gVAS, two transformations were applied to gLMS data, as shown in Figure 2. In the equal-interval transformed gLMS (eq-gLMS), each rating was transformed to preserve the relative distance between its two closest internal markings, as if it were marked on a scale with equally spaced labels, based on the methods of Green et al. (Green et al., 1993). For example, 15 on the gLMS is 9/11^{ths} between weak (6) and moderate (17), so 9/11^{ths} between weak (16.33) and moderate (50) on an equally spaced scale would be 47. In the "unlog" transformation, gLMS response values were plotted on a 1-100 logarithmic scale, then converted to the value that represented the distance from zero, as if it were rated using a standard 100mm scale ([$\log_{10}$(rating)]*50) (Figure 2). Normalized values were obtained by dividing participant responses by their rating for "the brightness of the sun on a clear day" then multiplying by 100. When applicable, normalization was performed before transformation.

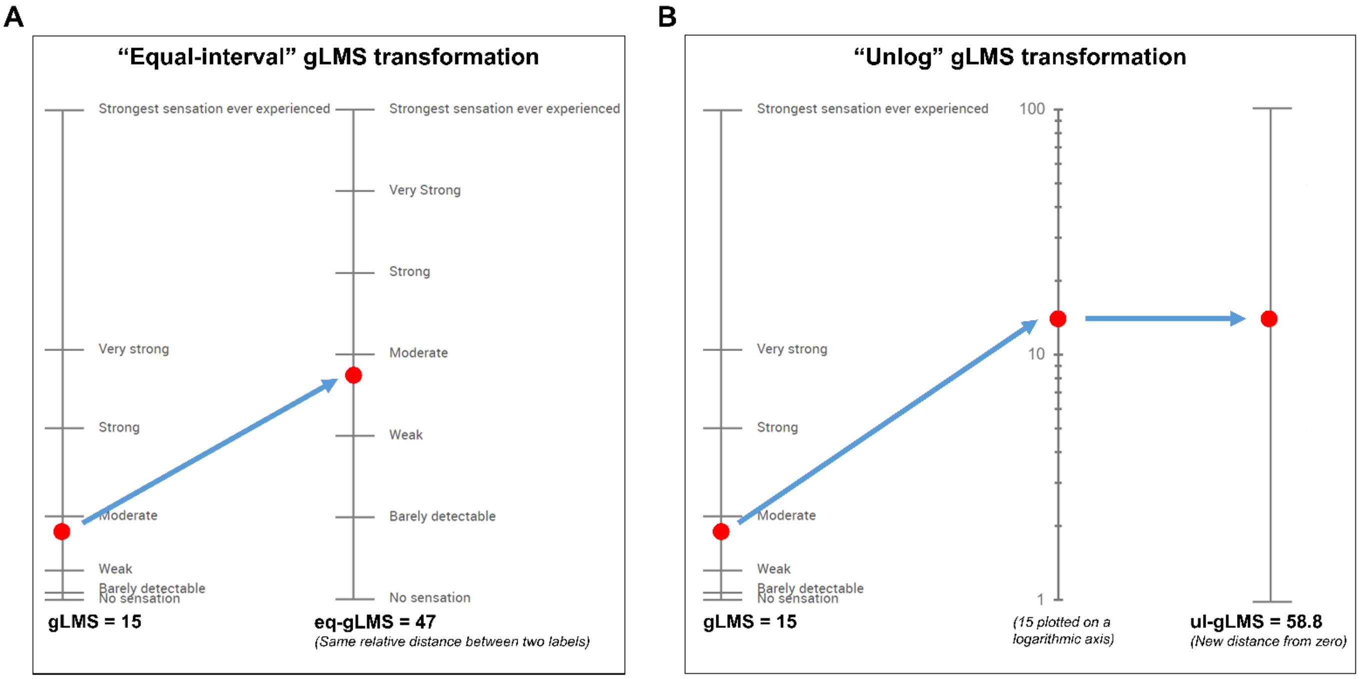


*Figure 2. Examples of transformations applied to data generated by the gLMS. A) In the equal-interval gLMS (eq-gLMS) transformation, 15 becomes 47, as both are 9/11ths between weak an moderate on their respective scales. B) In the unlog gLMS (ul-gLMS) transformation, 15 becomes 58.8 by plotting it on a 1-100 logarithmic scale and then converting it to the value that represents the distance from zero as if it were a 100mm scale ([$\log_{10}$(15)]*50).*

2.3 Statistical analysis

Data from participants that incorrectly used the scale (i.e., rated "the brightness of this room" higher than "the brightness of the sun on a clear day" or "the loudness of a whisper" higher than "the loudness of a shout") were excluded from statistical analysis (final n=183, 120 women, 63 men, 0 other, ages 8-80, mean age 38.5). The number of analyzed participants (and fails) for gLMS, gVAS, glVAS, and gwVAS, respectively was 48(2), 48(2), 46(4), and 43(4). Differences in sample means, resolving power, residual distributions, overall sample distribution, and categorical behavior for both cheese odor intensity and warm-up questions were compared as follows: 1) gLMS vs. gVAS; 2) gVAS vs. gLMS transformation/normalization to assess strategies for inter-scale comparisons; and 3) gVAS, glVAS, gwVAS, and eq-gLMS to assess the impact of label presence and spacing. Statistical significance was set at $p<0.05$, with no adjustments for multiple comparisons in order to set stricter standards for accepting the null hypothesis of no differences between scales.

Differences in sample means were assessed using linear mixed models ("proc mixed") in SAS 9.4. Participant was set as the repeated factor using the autoregressive covariance structure (data sorted by cheese type, then scale, participant ID, and cheese testing order) and the Kenward-Roger approximation for denominator degrees of freedom. Pair-wise comparisons within a cheese type (or question, in the case of analyzing the warm-up data) were assessed using the LSmeans function. To observe how scale type influenced resolving power, sample means were compared using a similar linear mixed model, with the exception that data was analyzed by scale rather than by cheese (following a sorting first by scale, then cheese). Residuals plots for each scale type were assessed qualitatively based on SAS residual outputs. Differences in response distributions between scales were determined using the Kolmogorov-Smirnov test in R-Studio, R version 3.4.3.

Categorical behavior was defined as ratings falling within 1 point of the scale label (Hayes et al., 2013; Lawless, Popper, & Kroll, 2010). Values for internal labels were rounded to the nearest whole number, as half points were not recorded. A Fisher's Exact test comparing the proportion of categorical behavior to the proportion that would occur by chance alone (0.19) was used to determine the significance of categorical behavior. To visualize categorical behavior, participant responses were assigned to the label category it most closely approximated (Schutz & Cardello, 2001).

3. Results

3.1 gLMS and gVAS comparison and the effects of data approximation strategies on sample means, response distribution, resolving power, and residuals.

Consistent with other reports (Hayes et al., 2013), we observed significantly lower mean odor intensity ratings generated by the gLMS compared with the gVAS (Figure 3, additional details found in Supplemental table 1). Overall response distribution was also significantly different between the gVAS and gLMS for all samples (Figure 4). Similar trends were observed in the warm-up data (Supplemental figure 1, Supplemental table 2). Sample rank order and discrimination sensitivity did not noticeably change between the two scales (Supplemental table 1), consistent with previous reports (Hayes et al., 2013). Residuals from the gLMS data showed greater positive skew than those from the gVAS (Figure 5).

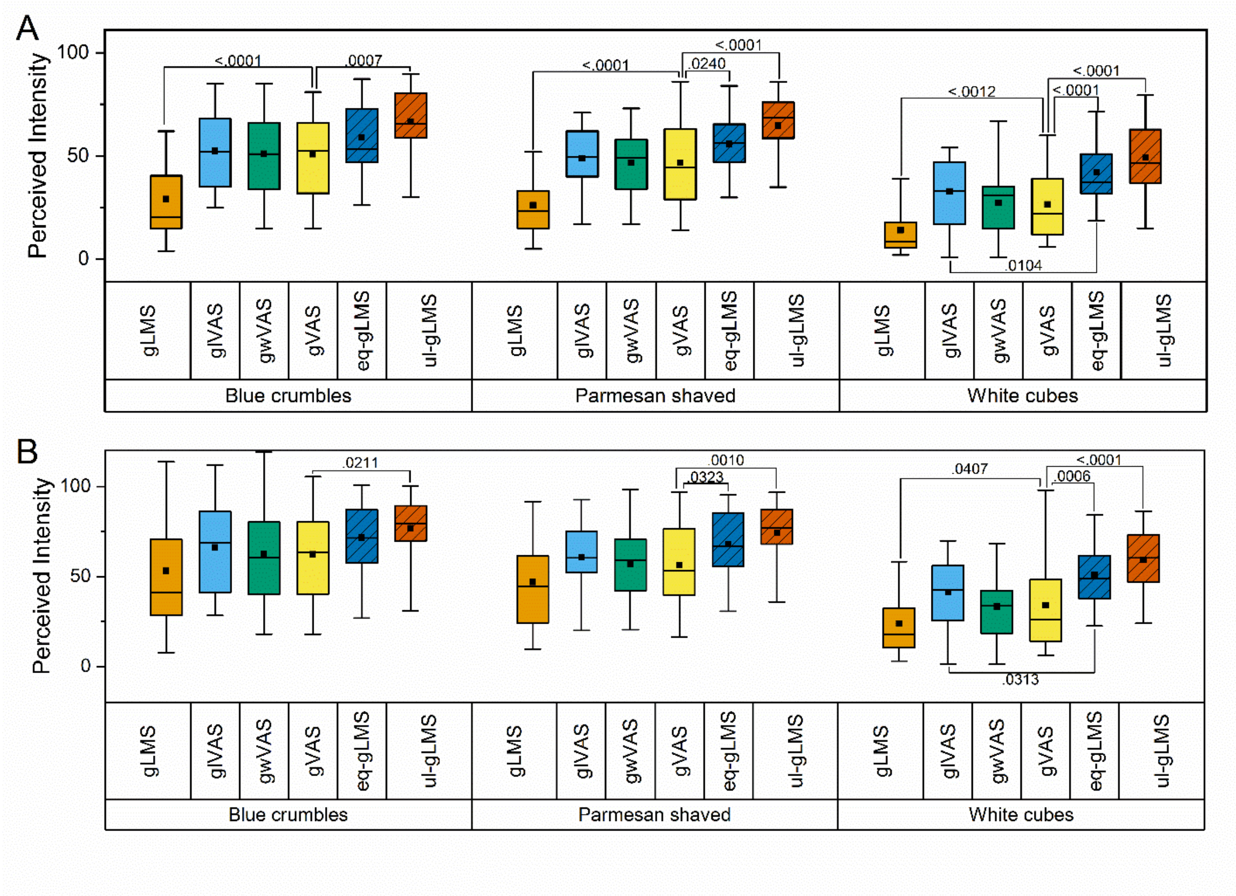


Figure 3. Box plots of raw a) and normalized b) cheese odor. Two gLMS transformations (equally spaced-gLMS and unLog-gLMS, indicated by hashed fill pattern) are also included. Boxes enclose the middle 50% of responses; the median is indicated by the center line. Scale means are represented by a square. Whiskers show 5^{th} and 95^{th} percentiles. Relevant pair-wise comparisons where $p < 0.05$ are displayed; no adjustments were made for multiple comparisons. Cheese odor differences within a scale are displayed in Supplemental table 1.

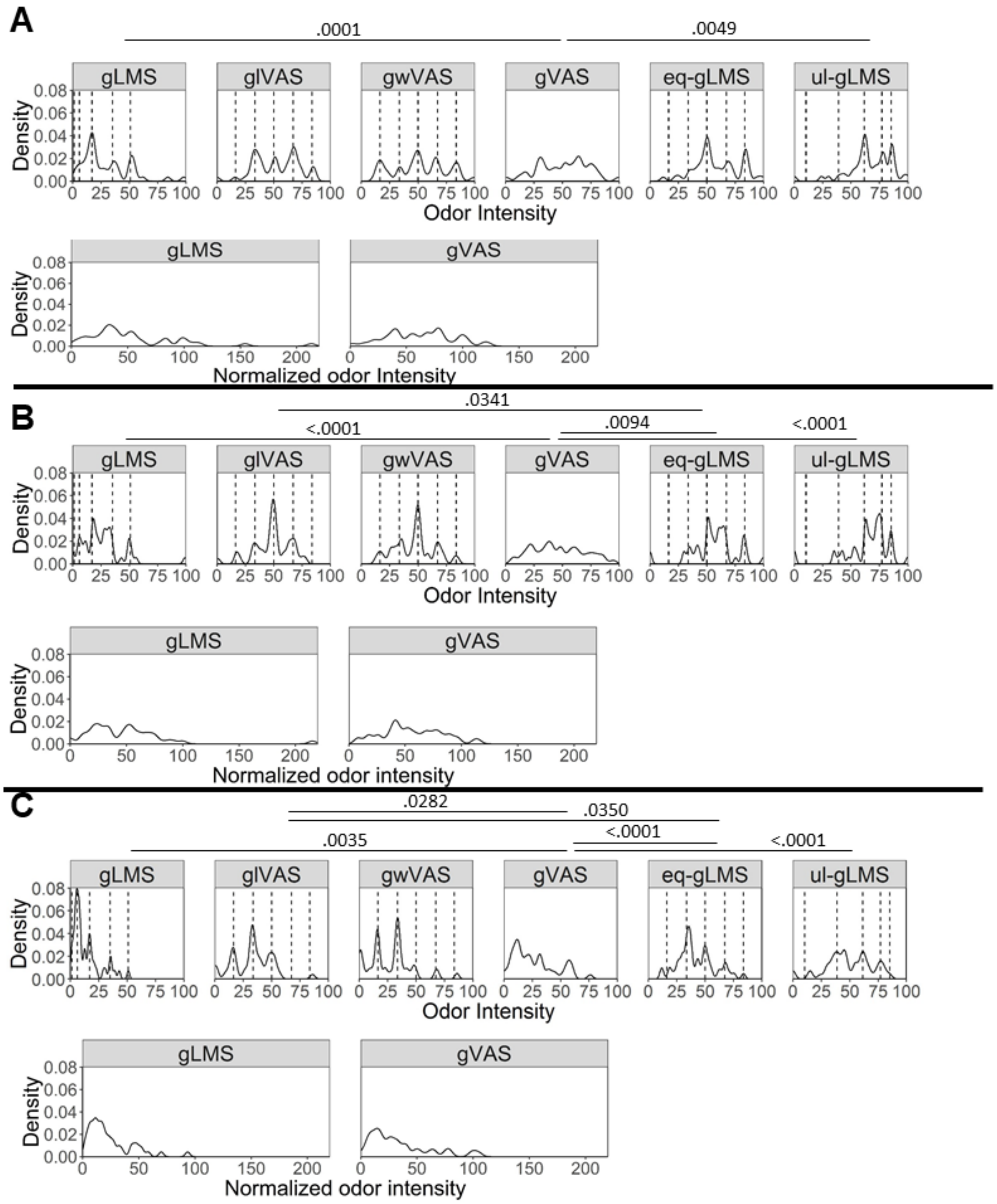


Figure 4. Kernel density estimates for cheese odor from raw data, transformed gLMS data and selected normalized responses for a) blue, b) parmesan, and c) white cheese. Dashed lines indicate the location of internal labels, when present. Comparisons within non-normalized data where $p < 0.05$ using the Kolmogorov-Smirnov test are displayed; no adjustments were made for multiple comparisons. Kolmogorov-Smirnov tests between normalized gLMS vs. gVAS were greater than 0.05 for all chese types.

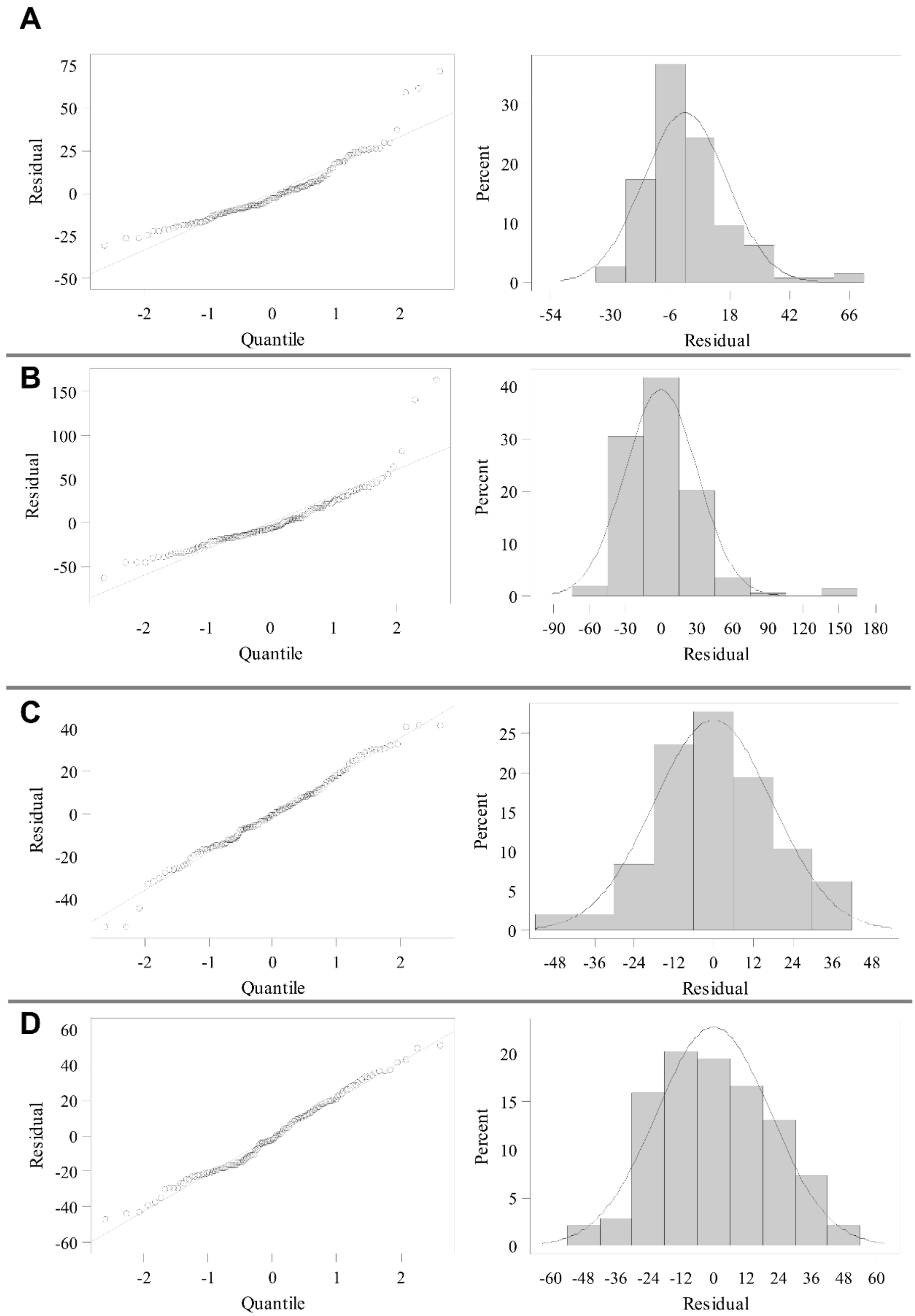


Figure 5 Residual plots of a) raw, b) normalized, and c) equal-interval-transformed gLMS and d) gVAS data, as generated using PROC MIXED in SAS 9.4. Data from all cheeses within a scale were combined to generate a single set of plots.

To investigate gVAS and gLMS data approximation strategies that could facilitate inter-scale comparisons, we next compared gVAS means with transformed and/or normalized gLMS values (Figures 3 and 4); fewer differences could indicate an appropriate comparison. Following the equal-interval transformation, sample means and response distribution between the gLMS and gVAS were significantly different in two of three cheese samples. Following the "unlog" transformation, response distribution and sample means were still different for all cheese types. Although fewer differences between the gLMS and gVAS data were observed following normalization, in line with findings by others (Hayes et al., 2013), higher variance was also observed; thus, statistical power to find differences was also reduced. Transforming normalized data increased the number of differences between the gLMS and gVAS compared to non-transformed normalized data. Normalization did not alter conclusions in warm-up data; mean responses and response distributions between gVAS and gLMS were still different. The equal-interval transformation, more than normalization, reduced the skewness of gLMS residuals (Figure 5).

3.2 Effect of scale elements on sample means, response distribution, resolving power, and categorical behavior.

To evaluate the effect of label spacing on participant behavior, we compared the glVAS with the eq-gLMS transformation. Consistent with findings by Green et. al (Green et al., 1993), we observed a tendency for higher sample means in eq-gLMS responses compared with glVAS responses (Figure 3). However, this difference was only statistically significant for the cheese with the lowest odor intensity (white cheese). Likewise, differences in mean warm-up responses between the eq-gLMS and glVAS were only observed for "the loudness of a whisper."

Normalization did not alter conclusions regarding sample mean comparison of these two scales. Differences between the response distribution of the eq-gLMS and glVAS reached statistical significance for two of three cheese samples (Figure 4). Relative to the eq-gLMS, participants using the glVAS showed a greater use of the lower end of the scale, as illustrated by shifted kernel density estimates (Figure 4) and redisplayed as a categorical data visualization for added clarity (Figure 6).

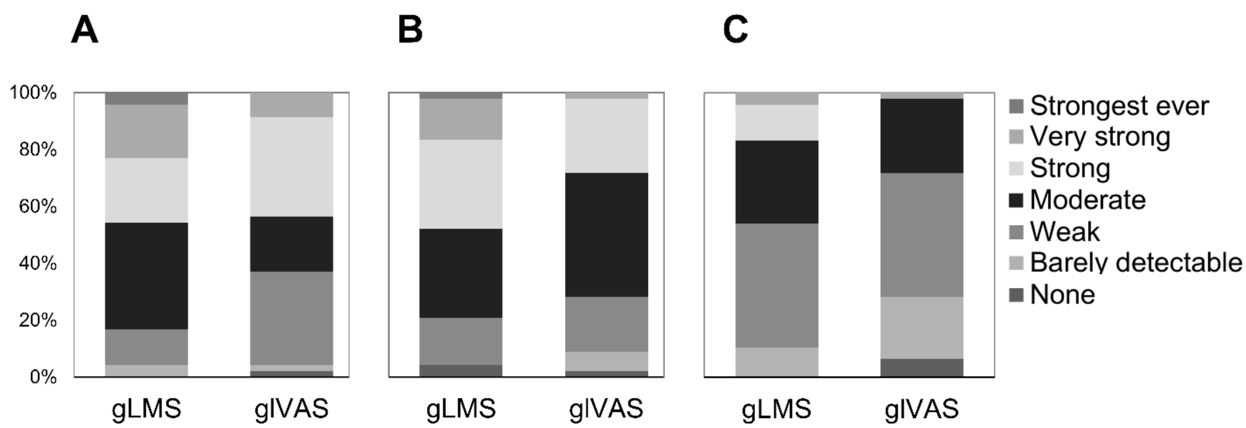


Figure 6. Visualization of "categories" following conversion of raw data to a categorical scale for a) blue, b) parmesan, and c) white cheese odor.

We next investigated the effect of equally-spaced internal labels with (glVAS) or without (gwVAS) lines on participant responses compared with no labels (gVAS). Both labeling approaches had a minimal effect on mean intensity rating and response distribution; only the response distribution of white cheese was statistically significant between the gVAS and the glVAS (Figures 3 and 4). Similarly, minimal differences were observed when labels were added to the warm-up questions (Supplemental figure 1).

Consistent with others' observations (Hayes et al., 2013), we observed categorical behavior in gLMS responses for cheese odor intensity and warm-up questions (qualitatively demonstrated by kernel density estimates in Figure 4). Indeed, 50% of gLMS responses fell within 1 point of the labeled lines (Table 2), which accounts for only 19% of the scale. The presence of words, with or without tick marks, was sufficient to induce categorical behavior, as categorical behavior was observed for both the glVAS and gwVAS.

Table 2. Categorical behavior (defined as marking within one point of an internal label) for each scale for both a) cheese odor and b) warm-up questions. A Fisher's Exact test comparing the proportion of categorical behavior to the proportion that would occur by chance alone (0.19) was used to determine the significance of categorical behavior.

A	gLMS	glVAS	gwVAS
Categorical marks	72	53	61
Total responses	144	138	129
% Categorical	50.0	38.4	47.3
Lower Wald's CI	41.8	30.3	38.7
High Wald's CI	58.2	46.5	55.9
Test statistic[1]	<.0001	<.0001	<.0001

[1]Fisher's exact test vs. 0.19, two-sided

B	gLMS	glVAS	gwVAS
Categorical marks	134	116	105
Total responses	276	276	258
% Categorical	48.6	42.0	40.7
Lower Wald's CI	0.4265	0.3621	0.347
High Wald's CI	0.5445	0.4785	0.4669
Test statistic[1]	<.0001	<.0001	<.0001

[1]Fisher's exact test vs. 0.19, two-sided

4. Discussion

In this study, we systematically compared strategies to approximate gLMS and gVAS data to explore the practicality of inter-scale comparisons. Our findings provide limited support for the use of cross-modal normalization or transformation in this context. Furthermore, we explored the effect of scale elements (i.e., internal label presence and spacing) on sample means, response distribution (including categorical behavior), resolving power, and residual distribution. We observed that for lower intensity samples, participants marked closer to more intense descriptors on the gLMS (which compresses the descriptors on the lower end of the scale) compared to a scale where those same descriptors were equally spaced. Additionally, the presence of labels, with or without lines, induced categorical behavior.

4.1 gLMS vs. gVAS

The first objective of this study was to compare two widely used scales in sensory analysis, the gLMS and gVAS, and whether transformation and/or normalization facilitated their comparison. Our observation of lower samples means in the gLMS is consistent with previous observations that compression of gLMS sample means is due in part to internal labeling (Hayes et al., 2013), as anchors are identical between these two scales. Among our limited sample set, we detected no differences in resolving power between the gLMS and gVAS, consistent with other reports (Hayes et al., 2013). Although some have proposed that a gLMS may have reduced discrimination sensitivity due to response compression (Cardello, Lawless, & Schutz, 2008; Lawless, Popper, et al., 2010), scale compression may also increase resolving power due to compressed variance (Cardello et al., 2008). Our findings are consistent with others that have failed to detect differences in rank order and sensitivity among scales with differences in response compression (Cardello et al., 2008; Hayes et al., 2013; Lawless, Popper, et al., 2010; Ludy & Mattes, 2011). Although labeled magnitude scales may offer advantages when extreme ratings are expected (Bartoshuk et al., 2003; Cardello et al., 2008; Lim et al., 2009; Schutz & Cardello, 2001), for intensities encountered in everyday experiences, there may be no advantage of one scale over others (Lawless, Sinopoli, & Chapman, 2010; Ludy & Mattes, 2011). Clearly, testing context must be considered when selecting the most appropriate scale.

Parametric statistical tests, which are commonly used to analyze results from both the gLMS and gVAS, require an assumption of equal variances and normally distributed residuals. However, these assumptions may be violated more often when using the gLMS, as we observed that residuals from the gVAS were less skewed than those from the gLMS. Additional investigation is needed to quantify the extent that violations of statistical assumptions influence conclusions from analyzing such data with parametric tests; however, from a technical perspective, researchers are advised to check the distributions of residuals if a gLMS is selected and transform the data as appropriate to correct the issue.

4.2 Data approximation strategies

Transformations and cross-modal normalizations of raw data have been performed as strategies to facilitate inter-scale comparisons (Hayes et al., 2013; Lim et al., 2009; Schutz & Cardello, 2001). In evaluating two transformation strategies for gVAS-gLMS comparisons, we observed that the equal-interval transformation eliminated statistical differences between sample means for only a limited number of samples; thus, our results do not support the use of either attempted transformation strategy for inter-scale comparisons. While normalizing responses resulted in fewer differences in sample means and response distributions between the gLMS and the gVAS, greater variance was also observed; thus we conclude that our failure to find differences was likely due to reduced statistical power rather than improved data approximation. Although some have proposed that normalizing raw data to a participant's own rating of a cross-modal standard may control for idiosyncratic or systematic differences in scale usage (Bartoshuk et al., 2004), more studies are needed to determine the most appropriate situations for the use of normalized data, and caution should be used when using normalized data to draw conclusions regarding sample differences.

4.3 Influence of label spacing and presence

Early psychophysicists noted that intensity labels such as "weak" and "strong" do not possess interval-level qualities (Borg, 1982; Lasagna, 1960). Furthermore, others have argued that spacing labels consistent with the relative strength of their perceptual intensity is important for proper scale usage (Green et al., 1993). Consequently, many researchers have attempted to quantitatively determine appropriate spacing of semantic descriptors to generate ratio-level level, including those that developed labeled magnitude scales (Green et al., 1993; Moskowitz, 1977; Schutz & Cardello, 2001). In the present work, we revisited the effect of label spacing by comparing responses from a scale with equally spaced labels (glVAS) and an equal-interval transformation of semi-logarithmically spaced labels (eq-gLMS). As eq-gLMS sample means were significantly higher than glVAS means only for weaker modalities (odor intensity of white cheese, loudness of a whisper), we suspect that end-use avoidance (Lawless & Heymann, 2010) influenced participant behavior. Similarly, Green and others observed higher responses from their LMS compared to a transformed equal-interval scale; although, in contrast to our findings, statistical significance was only observed for higher intensity samples (Green et al., 1993). As the gLMS was designed in part to accommodate “ceiling” effects (L. M. Bartoshuk et al., 2004), we suggest that the gLMS may not be appropriate for the evaluation of low-intensity samples by untrained participants.

In addition to label spacing, we also considered the effect of label presence on categorical behavior, or the tendency of participants to mark primarily where labels occur (Cardello et al., 2008; Hayes et al., 2013; Lawless, Popper, et al., 2010). We observed categorical behavior in all scales containing semantic labels, independent of internal line presence. If participants make category rather than ratio-level judgements, label spacing becomes less relevant because the scale no longer truly possesses continuous or ratio properties (Lawless, 2013). When participants make category judgements, the issue of end-use avoidance becomes more salient, as end-use avoidance is higher in category scales than continuous scales (Schutz & Cardello, 2001). We note that labels may not be necessary to generate ratio-level data, as ratio-level data has been generated using the gVAS (Hayes et al., 2013). Taken together, these observations further support the exercise of caution when using the gLMS for assessment of low-intensity samples.

Despite categorical behavior, we observed similar sample means, response distributions, and resolving power between the label-less gVAS and the labeled glVAS and gwVAS. Together, this suggests a minimal impact of categorical behavior on study conclusions, at least in our sample set. Although the presence of labels may influence behaviors that violate assumptions of truly normally distributed responses, whether this categorical behavior actually alters conclusions from data analysis is still unknown (Hayes et al., 2013).

Although labels can induce categorical behavior, internal markings may also make scales easier to use and provide meaningful qualitative information (Hayes et al., 2013). However, qualitative semantic labels also possess inherent limitations, as the same descriptors can have different meanings to different people (Bartoshuk et al., 2005; Bartoshuk et al., 2003). Thus, when selecting the appropriate scale for sensory research, the researcher should consider the value of qualitative information, expected sample intensity, and potential effects of categorical behavior on data analysis.

Conclusions from the current study must be interpreted within context of several limitations. First, each participant only used one scale to evaluate cheese odor intensity rather than using all scales. Additionally, only the odor of three quite distinct cheeses were assessed, so results may not apply to alternative sample types or a more internally similar sample set. Although the somewhat noisy testing environment of a fair booth is a study limitation, it also reflects a more realistic setting for how people actually consume food.

5. Conclusion

After exploring methods to compare data generated by the gLMS and gVAS, we found limited evidence to support transformation or cross-modal normalization as appropriate data approximation strategies. Although normalization may provide a very rough approximation to compare response distributions, the greater variance suggests caution should be used when drawing conclusions regarding resolving power. Despite identical scale anchors, differences in internal labels influence participant response behavior and thus hinder inter-scale comparisons. These data support previous recommendations that the gLMS and gVAS are not one-to-one comparable (Hayes et al., 2013), even when transformed or normalized. Our investigation of label spacing and presence revealed that words alone are sufficient to induce categorical behavior, and suggested that compressed labels (as in the gLMS) may cause participants to mark near higher intensity descriptors when low-intensity samples are evaluated. Consistent with several other reports, we did not find any systematic advantage of one scale over another (Cardello et al., 2008; Hayes et al., 2013; Lawless, Popper, et al., 2010; Ludy & Mattes, 2011). Overall, we conclude that all of these scales are valid and useful tools, but we urge researchers to carefully select the scale and/or transformation method best suited for their samples, participant group, and data analysis approach.

Acknowledgments

The authors thank the American Dairy Association of Indiana for providing the cheese samples and Miguel Odron for his assistance in carrying out the study. The authors declare no conflicts of interest.

References

Bartoshuk, L.M., Fast, K., & Snyder, D. J. (2005). Differences in our sensory worlds: Invalid comparisons with labeled scales. *Current Directions in Psychological Science*, *14*(3), 122–125. https://doi.org/10.1111/j.0963-7214.2005.00346.x

Bartoshuk, L. M., Duffy, V. B., Fast, K., Green, B. G., Prutkin, J., & Snyder, D. J. (2003). Labeled scales (e.g., category, Likert, VAS) and invalid across-group comparisons: What we have learned from genetic variation in taste. *Food Quality and Preference*, *14*(2), 125–138. https://doi.org/10.1016/S0950-3293(02)00077-0

Bartoshuk, L. M., Duffy, V. B., Green, B. G., Hoffman, H. J., Ko, C. W., Lucchina, L. A., … Weiffenbach, J. M. (2004). Valid across-group comparisons with labeled scales: the gLMS versus magnitude matching. *Physiology and Behavior*, *82*(1), 109–114. https://doi.org/10.1016/j.physbeh.2004.02.033

Borg, G. (1982). A category scale with ratio properties for intermodal and interindividual comparisons (pp. 25–34). Presented at the XXII International Congress of Psychology, North-Holland Pub. Co.

Cardello, A., Lawless, H. T., & Schutz, H. G. (2008). Effects of extreme anchors and interior label spacing on labeled affective magnitude scales. *Food Quality and Preference*, *19*(5), 473–480. https://doi.org/10.1016/j.foodqual.2008.02.003

Duffy, V. B., Peterson, J. M., & Bartoshuk, L. M. (2004). Associations between taste genetics, oral sensation and alcohol intake. *Physiology and Behavior*, *82*(2–3), 435–445. https://doi.org/10.1016/j.physbeh.2004.04.060

Green, B. G., Shaffer, G. S., & Gilmore, M. M. (1993). Derivation and evaluation of a semantic scale of oral sensation magnitude with apparent ratio properties. *Chemical Senses*, *18*(6), 683–702. https://doi.org/10.1093/chemse/18.6.683

Hayes, J. E., Allen, A. L., & Bennett, S. M. (2013). Direct comparison of the generalized visual analog scale (gVAS) and general labeled magnitude scale (gLMS). *Food Quality and Preference*, *28*(1), 36–44. https://doi.org/10.1016/j.foodqual.2012.07.012

Kalva, J. J., Sims, C. A., Puentes, L. A., Snyder, D. J., & Bartoshuk, L. M. (2014). Comparison of the hedonic general labeled magnitude scale with the hedonic 9-point scale. *Journal of Food Science*, *79*(2), S238–S245. https://doi.org/10.1111/1750-3841.12342

Lasagna, L. (1960). The clinical measurement of pain. *Annals of the New York Academy of Sciences*, *86*(1), 28–37. https://doi.org/10.1111/j.1749-6632.1960.tb42788.x

Lawless, H. T. (2013). *Quantitative Sensory Analysis*. West Sussex, UK: Wiley-Blackwell.

Lawless, H. T., & Heymann, H. (2010). *Sensory evaluation of food: principles and practices*. Springer Science & Business Media.

Lawless, H. T., Popper, R., & Kroll, B. J. (2010). A comparison of the labeled magnitude (LAM) scale, an 11-point category scale and the traditional 9-point hedonic scale. *Food Quality and Preference*, *21*(1), 4–12. https://doi.org/10.1016/j.foodqual.2009.06.009

Lawless, H. T., Sinopoli, D., & Chapman, K. W. (2010). A comparison of the labeled affective magnitude scale and the 9-point hedonic scale and examination of categorical behavior. *Journal of Sensory Studies*, *25*(s1), 54–66. https://doi.org/10.1111/j.1745-459X.2010.00279.x

Lim, J., Wood, A., & Green, B. G. (2009). Derivation and evaluation of a labeled hedonic scale. *Chemical Senses*, *34*, *34*(9, 9), 739, 739–751. https://doi.org/10.1093/chemse/bjp054, 10.1093/chemse/bjp054

Ludy, M.-J., & Mattes, R. D. (2011). Noxious stimuli sensitivity in regular spicy food users and non-users: Comparison of visual analog and general labeled magnitude scaling. *Chemosensory Perception*, *4*(4), 123–133. https://doi.org/10.1007/s12078-011-9100-x

Moskowitz, H. R. (1977). Magnitude estimation: Notes on what, how, when, and why to use it. *Journal of Food Quality*, *1*(3), 195–227. https://doi.org/10.1111/j.1745-4557.1977.tb00942.x

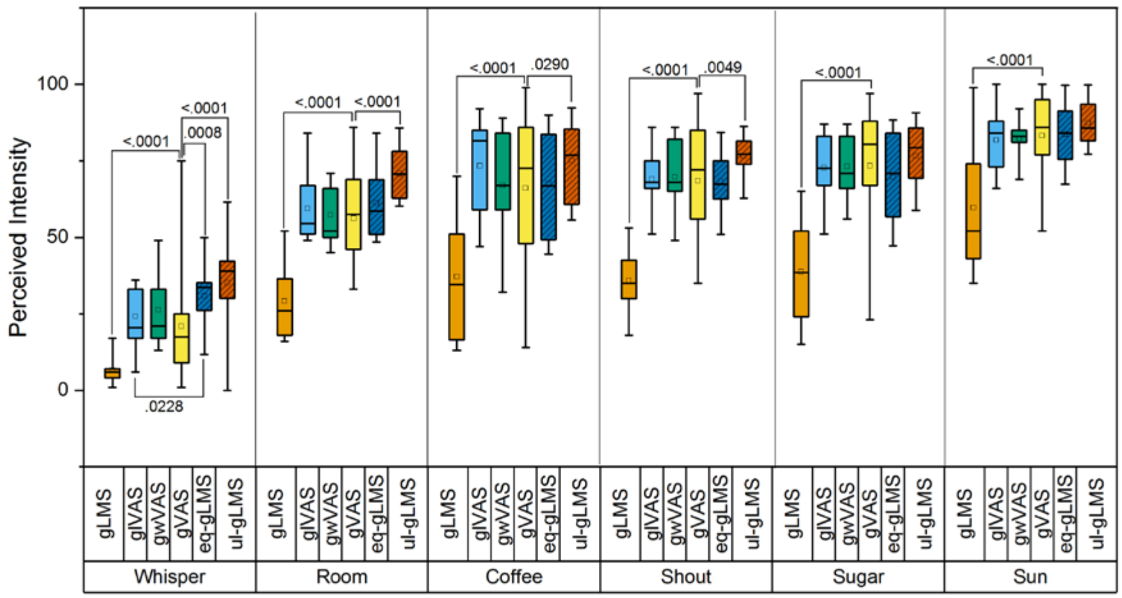
Price, D. D., McGrath, P. A., Rafii, A., & Buckingham, B. (1983). The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. *Pain*, *17*(1), 45–56.

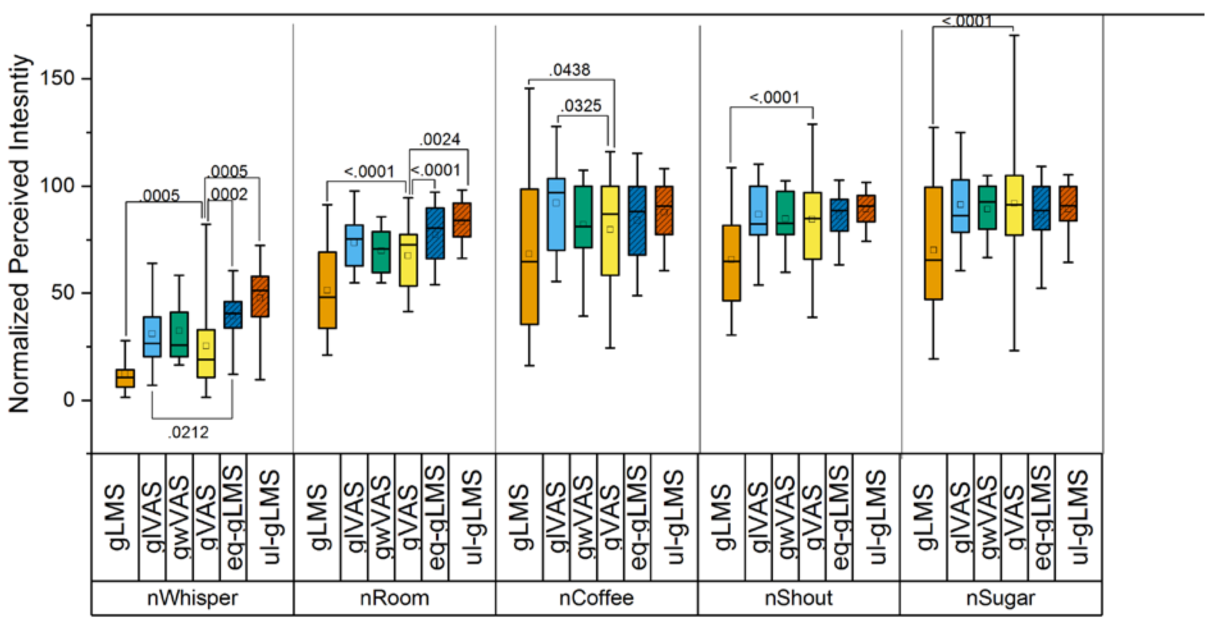
Schutz, H. G., & Cardello, A. V. (2001). A labeled affective magnitude (LAM) scale for assessing food liking/dislking. *Journal of Sensory Studies*, *16*(2), 117–159. https://doi.org/10.1111/j.1745-459X.2001.tb00293.x

Webb, J., Bolhuis, D. P., Cicerale, S., Hayes, J. E., & Keast, R. (2015). The relationships between common measurements of taste function. *Chemosensory Perception, 8*(1), 11–18. https://doi.org/10.1007/s12078-015-9183-x

Zealley, A. K., & Aitken, R. C. (1969). Measurement of mood. *Proceedings of the Royal Society of Medicine*, *62*(10), 993–996.

Supplemental files





Supplemental figure 1. Box plots of raw a) and normalized b) warm-up data. Two gLMS transformations (equally spaced-gLMS and unLog-gLMS, indicated by hashed fill pattern) are also included. Boxes enclose the middle 50% of responses; the median is indicated by the center line. Scale means are represented by a square. Whiskers show 5th and 95th percentiles. Relevant pair-wise comparisons were $p < 0.05$ are displayed; no adjustments were made for multiple comparisons. Question differences within a scale are displayed in table 1.

Supplemental table 1. Differences between response means within a scale (displayed in figure and supplemental figure 1). Normalized values are indicated by “n”. Samples with different superscripts indicate a p-value < 0.05 using the least squares difference between two samples/questions within a scale; no adjustments were made for multiple comparisons.

	gLMS	glVAS	gwVAS	gVAS	eq-gLMS	unlog-gLMS
Blue	29.0^{a}	52.8^{a}	50.8^{a}	50.8^{a}	58.7^{a}	66.5^{a}
Parmesan	26.1^{a}	48.9^{a}	45.8^{a}	46.6^{a}	55.7^{a}	64.5^{a}
White	14.1^{b}	32.4^{b}	27.8^{b}	26.8^{b}	42.3^{b}	49.4^{b}
nBlue	48.3^{a}	66.9^{a}	61.9^{a}	62.1^{a}	99.6^{a}	113.8^{a}
nParmesan	41.7^{a}	60.6^{a}	55.8^{a}	56.4^{a}	90.2^{a}	105.9^{a}
nWhite	21.9^{b}	40.6^{b}	33.9^{b}	34.7^{b}	68.1^{b}	79.4^{b}
Whisper	6.7^{a}	24.2^{a}	26.3^{a}	20.9^{a}	30.8^{a}	35.2^{a}
Room	29.2^{b}	59.5^{b}	57.3^{b}	56.3^{b}	61.2^{b}	70.8^{b}
Coffee	*37.2*^{bc}	*73.4*^{c}	66.9^{c}	66.1^{bc}	67.5^{bc}	75.2^{b}
Shout	*35.9*^{bc}	*69.0*^{c}	69.8^{c}	68.5^{c}	68.2^{bc}	*76.7*^{b}
Sugar	38.8^{c}	*72.7*^{c}	73.1^{c}	73.4^{cd}	69.8^{c}	*76.6*^{b}
Sun	59.8^{d}	81.7^{d}	82.4^{d}	83.2^{d}	83.8^{d}	87.5^{c}
nWhisper	12.4^{a}	31.1^{a}	32.3^{a}	25.4^{a}	39.5^{a}	47.7^{a}
nRoom	51.4^{b}	73.6^{b}	69.8^{b}	67.4^{b}	77.4^{b}	83.4^{b}
nCoffee	*68.2*^{c}	*92.1*^{cd}	82.0^{c}	79.7^{bc}	84.0^{bc}	87.8^{b}
nShout	*65.7*^{c}	*86.8*^{c}	84.8^{c}	84.5^{c}	86.1^{c}	*89.3*^{b}
nSugar	70.1^{c}	*91.3*^{cd}	89.2^{c}	91.8^{cd}	86.6^{c}	*89.2*^{b}
nSun	100.0^{d}	100.0^{d}	100.0^{d}	100.0^{d}	100.0^{d}	100.0^{c}

Supplemental table 2. Differences in response distributions of raw, transformed, and normalized (n, gLMS and gVAS only) warm-up questions, as measured using the Kolmogorov-Smirnov test. No adjustments were made for multiple comparisons.

Question	Comparison		KS statistic
Whisper	gLMS	gVAS	<.0001
Whisper	glVAS	gVAS	0.0037
Whisper	gwVAS	gVAS	0.0049
Whisper	glVAS	gwVAS	0.9123
Whisper	eq-gLMS	gVAS	0.0000
Whisper	ul-gLMS	gVAS	0.0000
Whisper	glVAS	eq-gLMS	0.0036
Whisper	n gLMS	n gVAS	0.0002
Room	gLMS	gVAS	<.0001
Room	glVAS	gVAS	0.0874
Room	gwVAS	gVAS	0.2515
Room	glVAS	gwVAS	0.8571
Room	eq-gLMS	gVAS	0.1308
Room	ul-gLMS	gVAS	<.0001
Room	glVAS	eq-gLMS	0.5892
Room	n gLMS	n gVAS	0.0006
Coffee	gLMS	gVAS	<.0001
Coffee	glVAS	gVAS	0.2271
Coffee	gwVAS	gVAS	0.5702
Coffee	glVAS	gwVAS	0.2715
Coffee	eq-gLMS	gVAS	0.4695
Coffee	ul-gLMS	gVAS	0.0469
Coffee	glVAS	eq-gLMS	0.1447
Coffee	n gLMS	n gVAS	0.0155
Shout	gLMS	gVAS	<.0001
Shout	glVAS	gVAS	0.2271
Shout	gwVAS	gVAS	0.6179
Shout	glVAS	gwVAS	0.3472
Shout	eq-gLMS	gVAS	0.1335
Shout	ul-gLMS	gVAS	0.0033
Shout	glVAS	eq-gLMS	0.6408
Shout	n gLMS	n gVAS	0.0021
Sugar	gLMS	gVAS	<.0001
Sugar	glVAS	gVAS	0.2271
Sugar	gwVAS	gVAS	0.1625
Sugar	glVAS	gwVAS	0.9894
Sugar	eq-gLMS	gVAS	0.1976
Sugar	ul-gLMS	gVAS	0.6260
Sugar	glVAS	eq-gLMS	0.2090
Sugar	n gLMS	n gVAS	0.0003
Sun	gLMS	gVAS	<.0001
Sun	glVAS	gVAS	0.1440
Sun	gwVAS	gVAS	0.0134
Sun	glVAS	gwVAS	0.3063
Sun	eq-gLMS	gVAS	0.5530
Sun	ul-gLMS	gVAS	0.1308
Sun	glVAS	eq-gLMS	0.7648
Sun	n gLMS	n gVAS	--

Code

```
proc sort data=Cheese;
by cheese scale ID order;
run;
*Comparing SCALE by cheese odor;
ods graphics on;
title 'Scale differences by cheese odor';
ods output tests3=effects diffs=MeansDiffs LSmeans=LSmeans;
proc mixed data=Cheese;
     by cheese;
     class ID scale order;
     model Odor= scale order scale*order / residual s outp=pred;
     repeated / subject = id type= ar(1);
     lsmeans scale / diff adjust=tukey;
run;
proc print data=effects;
run;
proc print data=MeansDiffs;
run;
proc print data=LSmeans;
run;

proc sort data=Cheese;
by scale cheese ID order;
run;
*Comparing cheese odor by SCALE;
ods graphics on;
title 'Odor by scale';
ods output tests3=effects diffs=MeansDiffs LSmeans=LSmeans;
proc mixed data=Cheese;
     by scale;
     class ID Cheese order;
     model Odor= cheese order cheese*order / residual s
outp=pred;
     repeated / subject = id type= ar(1);
     lsmeans cheese / diff adjust=tukey;
run;
proc print data=effects;
run;
proc print data=MeansDiffs;
run;
proc print data=LSmeans;
run;
```

```
proc freq data= ms_elms order=data;
     tables scale*YN / fisher;
     exact fisher;
     weight freq;
run;

proc freq data= Categ order=data;
     by scale;
     tables YN / binomial(equiv p=.19 margin=.05);
     weight freq;
run;

```