COGNITION AND EMOTION
2008, 22 (1), 180–192

Psychology Press
Taylor & Francis Group

BRIEF REPORT

Addressing measurement limitations in affective rating scales: Development of an empirical valence scale

David A. Lishner
University of Wisconsin Oshkosh, Oshkosh, WI, USA

Amy B. Cooter and David H. Zald
Vanderbilt University, Nashville, TN, USA

Traditional self-report measures suffer from weaknesses in either the quantitative or qualitative assessment of subjective experience. Researchers interested in the subjective intensity of oral sensation have attempted to reduce these scale limitations by developing rating scales with empirically determined placement of verbal descriptors along a continuous visual analogue scale continuum. In the present research, a similar empirical approach to scale construction was adopted to develop a rating scale of emotional valence. The potential benefits of using an empirically derived valence scale and techniques for validating the scale are discussed.

Emotion consists of multiple facets, including behavioural, physiological, and experiential components (Izard, 1977). Whereas the behavioural and physiological components can be readily measured in a variety of ways, assessing the experiential component is more challenging. At this time, the only available means of doing so is to use self-report rating scales. In emotion research, the two primary types of rating scales utilised are visual analogue scales (VAS; Hayes & Patterson, 1921) and Likert scales (LS; Likert, 1932).

Correspondence should be addressed to: David A. Lishner, Department of Psychology, University of Wisconsin Oshkosh, 800 Algoma Blvd., Oshkosh, WI 54901, USA. E-mail: lishnerd@uwosh.edu

DL was supported by NIMH training grant #T32-MH18921 during data collection.

We thank Linda Bartoshuk who challenged us to incorporate more modern rating scale techniques in emotion research. We also thank Liz Abernathy for help in collecting pilot data and Sean Jewett for assistance with computer programming.

www.psypress.com/cogemotion DOI: 10.1080/02699930701319139

Strengths and weaknesses of visual analogue and Likert scales

VAS consist of two verbal anchor descriptors that are placed on either end of a continuous line. Respondents provide ratings on VAS by indicating the point on the line that best represents the intensity of their current psychological experience. In contrast to the continuous line of VAS, LS consist of numeric points arranged along a discrete continuum. In addition to labels placed at both anchor points, the intermediary numeric points are labelled with verbal descriptors of varying intensity. When using LS, respondents indicate the descriptor that best represents the intensity of their current psychological experience. In addition to traditional VAS or LS, many researchers prefer to use a discrete rating scale where only the anchor points are labelled and intermediate points are simply marked by numerals. For the purpose of the following discussion, such scales are considered discrete visual analogue scales.

Two quantitative strengths have been attributed to VAS. First, VAS allow respondents to choose from a large number of continuous rating values, which may make them more amenable to parametric statistical analyses. Second, some have suggested that VAS provide interval-like data. Interval-like data are ideal because they provide rank-order information about rating values and equal spacing exists between neighbouring values along the entire scale continuum. When VAS include an objective absolute zero point as an anchor, they also may produce ratio-like data. Unfortunately, demonstrating that a rating scale of subjective experience actually generates ratio or interval-like data proves challenging as no gold standard exists for evaluating such a claim. Consequently, it may be more appropriate to assume that VAS generate ordinal data rather than interval-like data.

Although a number of quantitative strengths have been attributed to VAS, the absence of intermediary descriptors leads to their major weakness. Because the continuum lacks verbal descriptors, it is impossible to ascertain the qualitative intensity that corresponds to an intermediary rating. This means that intermediary ratings made by different individuals are not readily comparable. Even though different respondents may report similar ratings, without shared qualitative reference points an observer has no way of knowing whether the similar ratings indeed reflect similar experiential intensities.

In contrast to VAS, LS allow for stronger conclusions about the qualitative differences among the intermediary ratings of different respondents, because each discrete numeric point is designated by a verbal descriptor of intensity. Unfortunately, even though distances between verbal descriptors on LS appear equidistant, the actual experiential distances may be quite heterogeneous. For example, when using LS that assess emotional experience, the typical respondent may perceive the experiential distance

between rating values of *slight* and *moderate* as subjectively smaller than the experiential distances between rating values of *moderate* and *strong* even though *slight* and *strong* are depicted as equidistant from *moderate*. Such heterogeneity means that one can only make rank-order inferences about the relative differences between two or more ratings (i.e., one rating is higher or lower than a second rating). Furthermore, in contrast to VAS, the discrete structure of LS results in the generation of less continuous data.

A second weakness with LS is that the number of verbal descriptors is determined heuristically. The number of descriptors on LS typically ranges from four to five. However, it is unclear whether such a range constitutes an ideal number of descriptors. If too many descriptors are placed along the scale continuum, then respondents are forced to choose among different descriptors that cannot be experientially discriminated, whereas use of too few descriptors will lead to poor detection of meaningful experiential differences.

Scale construction and ceiling effects

A common limitation with VAS and LS is failure to address the potential for floor or ceiling effects. Specifically, detection of extreme reactions is highly dependent upon the anchor descriptors. Use of anchor descriptors that signify a relative extreme (e.g., *extremely*) rather than an absolute extreme (e.g., *most imaginable*) limits the ability to assess the most intense reactions. Although such limitations may have no impact in situations where experiential intensity is modest, such limitations may prove highly problematic in situations where experiential intensity is extreme. For instance, when provided with an anchor descriptor of *strongest imaginable* in taste studies, many participants indicate ratings that exceed the relative term *very strong*, a term that is used often as an anchor descriptor for VAS and LS (Bartoshuk, Fast, & Snyder, 2005; Green et al., 1996; Green, Shaffer, & Gilmore, 1993).

Addressing scale weaknesses using empirical methods

Based on a long line of sensation research using variations of VAS and LS (Bartoshuk & Marks, 1986; Borg, 1982; Marks, Borg, & Westerlund, 1992; Marks, Stevens, & Bartoshuk, 1988; Poultan, 1968; Stevens, 1957, 1971), Green et al. (1993) argue that one can develop labelled VAS that maximise the strengths and minimise the weaknesses of VAS and LS by using an empirical approach to locate descriptors along a visual analogue continuum. They illustrated such an approach by creating the Labelled Magnitude Scale (LMS), which assesses the intensity of oral sensation. To construct the LMS, Green et al. instructed participants to use a numeric scale to rate imagined oral sensations of different intensity and sensory modality. Participants were then asked to rate descriptors of varying intensity using the same scale,

including anchor descriptors of absolute minimal and maximal sensory intensity. The mean descriptor ratings were then used to determine where along the scale continuum to place the descriptors. In subsequent studies, the LMS has proven superior to traditional LS at detecting individual differences in taste sensitivity (Green et al., 1993, 1996).

Application of the LMS to the affective domain

The perception or subjective experience of emotional valence is linked to a variety of important psychophysiological phenomena (Benning, Patrick, & Lang, 2004; Dimberg, 1997; Hu & Wan, 2003; Lang, 1995; Lang, Bradley, & Cuthbert, 1990, 1992) and theoretically relevant personality traits (Schwerdtfeger, 2003; Vanman, Dawson, & Brennan, 1998). Recently, Bartoshuk et al. (2005) modified the LMS to assess the valence (hedonic) ratings of food. They did this by making a bipolar version of the LMS with the anchor descriptors *strongest imaginable disliking of any kind*, *neutral*, and *strongest imaginable liking of any kind*. They argue that such modification leads to improved measurement of taste preference in women (Snyder, Duffy, Hoffman, Ko, & Bartoshuk, 2003). Although the modified LMS appears promising for measuring food preference, which is likely associated with oral sensation, it is unclear if the LMS would offer promise as a measure of hedonic emotional experience (despite Bartoshuk et al.'s suggestion that the LMS structure will generalise across experiential modalities). Indeed, part of Green et al.'s (1993) rationale for developing the LMS was concern that an already existing empirically spaced scale (the Category Ratio Scale; Borg, 1982), which was developed to measure perceived exertion, did not appear to perform as well when applied to the measurement of sensory intensity. Consequently, the purpose of the present research was to develop a rating scale of emotional valence using Green et al.'s (1993) empirical technique.

METHOD

Participants

Participants were 34 introductory psychology students (20 men, 14 women) from Vanderbilt University and 40 non-student adult participants (14 men, 26 women) solicited from the local Nashville, Tennessee, community. The community participants were recruited to ensure that the ratings would generalise beyond a student population. Each participant completed written informed consent approved by the Vanderbilt University IRB. Student participants received credit to fulfill a psychology research experience requirement and community participants received payment in return for participation.

Procedure

Participants were run individually in sessions that lasted approximately 30 minutes. After completing informed consent, participants were seated in front of a computer screen and told that the study was designed to develop a rating scale to assess emotionally pleasant or unpleasant feelings. Participants were told that in Part 1 of the study they would rate images and in Part 2 of the study they would rate verbal descriptions.

During Part 1 of the study, participants were informed that they would be presented with 40 images that they were to rate using a scale ranging from −100 to 0 to 100. They were instructed to use the scale in the following manner:

> When using the numeric scale think of 0 as indicating a "neutral" feeling that is neither pleasant nor unpleasant. Think of positive numbers as indicating feeling pleasant, with 100 indicating the "most pleasant imaginable". Similarly, think of negative numbers as indicating feeling unpleasant, with −100 indicating the "most unpleasant imaginable".

They were to rate the degree of their feeling by using a mouse cursor to click on the desired scale point. Finally, participants were instructed as follows: "When rating a given image, try to rate how pleasant or unpleasant it makes you feel relative to other pleasant or unpleasant feelings of ALL kinds".

Image ratings. After receiving the directions, participants were shown a series of 40 images from the International Affective Picture System (IAPS) (Lang, Bradley, & Cuthbert, 1995a). The 40 images depicted scenes that varied in the degree of emotional valence they evoke in typical observers. These images were selected based on normative IAPS ratings of valence provided by Lang, Bradley, and Cuthbert (1995b). The images, ranging from most unpleasant to neutral to most pleasant, were as follows: burn victims, injured or starving children, male aggression toward women, car accidents, garbage, roaches, angry adult faces, boys with unpleasant expressions, men with neutral expressions, mugs, attractive men, naked women, motorcyclists, food, skyline, erotic heterosexual couples, butterflies, romantic couples, fathers with their infants, and infant faces. Two similar images were selected for each of these 20 picture categories.

Participants were asked to rate the emotional images to familiarise them with the rating task and to induce a variety of pleasant and unpleasant emotional experiences to reflect on in Part 2 of the study. Images were presented in a random order for each participant. Beneath each image, the numeric scale was displayed and participants indicated their emotional

reactions to the image using the mouse. Once they indicated their reactions, the next image was displayed.

Verbal descriptor ratings. After rating the 40 images, participants were informed that in Part 2 of the study they would be rating 19 verbal descriptions of different degrees of emotional pleasantness and unpleasantness. It was explained that they would rate each description using the same numeric scale used in Part 1 of the study. Participants were further instructed as follows: "When rating a given verbal description try to consider the FULL range of emotional pleasant and unpleasant feelings one might experience. Try to use your reactions to the images you viewed in Part 1 as a helpful guideline for doing this".

The following verbal descriptors were included for participants to rate: *Neutral*, *most (un)pleasant imaginable*, *barely (un)pleasant*, *a little (un)pleasant*, *slightly (un)pleasant*, *mildly (un)pleasant*, *moderately (un)pleasant*, *strongly (un)pleasant*, *very (un)pleasant*, and *extremely (un)pleasant*. Verbal descriptors were presented on a black screen with the numeric rating scale beneath them. As with the image ratings, once participants had indicated their reactions the next verbal descriptor was presented. The first three verbal descriptors were always the three anchor points (*neutral*, *most (un)pleasant imaginable*) presented in a random order. After participants rated these three anchor verbal descriptors, they then rated the remaining 16 descriptors, which were also presented in a random order.

RESULTS

Data screening

Prior to computing means for the descriptor ratings, responses for which it was clear that the participant had misread a given verbal descriptor were eliminated. For example, in a few cases participants rated a pleasant verbal descriptor as unpleasant or vice versa. Ratings were collapsed across both participant samples to provide more representative estimates of the descriptor means.

When constructing the LMS, Green et al. (1993) found that verbal descriptor ratings appeared to be log-normally distributed, so they transformed their rating data into logarithms prior to calculating rating scale means. Unlike Green et al., the majority of the present rating distributions appeared symmetrical and unimodal. The remaining distributions failed to show any systematic distribution pattern and the mean for each distribution seemed to reasonably capture its central tendency. Consequently, no transformations were applied to the rating data. Instead, any descriptor ratings that were more than two standard deviations away

from the mean rating of the corresponding descriptor were eliminated. Use of a liberal value of 2 standard deviations for eliminating outliers ensured that rating estimates more closely reflected normative perceptions of qualitative intensity. This data screening process resulted in 65 to 72 ratings per descriptor out of a total possible 74 ratings.

Scale construction

After eliminating outliers, mean descriptor ratings were recalculated. The means and 95% confidence intervals for each of the descriptors are presented in Table 1. Two approaches were used to determine whether participants could discriminate between neighbouring descriptors. First, repeated-measures *t*-tests were conducted to determine whether neighbouring means were significantly different from one another. Because multiple comparisons were made, a relatively conservative alpha level was used for establishing statistical significance ($\alpha = .001$). Second, discrimination between neighbouring descriptors was assumed if mean difference effect sizes were large (Cohen's $d > .80$).

All neighbouring means produced large mean difference effect sizes and were significantly different from one another except for the following pairs:

TABLE 1
Means and confidence intervals for the verbal descriptor ratings

Verbal descriptor	*Mean*	*95% Confidence Interval*
Most unpleasant imaginable	−99.32 (66)	−98.77, −99.87
Extremely unpleasant	−84.80 (69)	−82.10, −87.50
Strongly unpleasant	−71.22 (68)	−67.57, −74.87
Very unpleasant	−68.91 (68)	−65.68, −72.15
Moderately unpleasant	−37.43 (70)	−34.42, −40.44
Mildly unpleasant	−24.29 (69)	−21.75, −26.83
Slightly unpleasant	−11.46 (67)	−10.10, −12.82
A little unpleasant	−11.25 (67)	−9.86, −12.65
Barely unpleasant	−6.10 (68)	−5.22, −6.98
Neutral	−0.50 (72)	−0.21, −0.79
Barely pleasant	5.90 (69)	5.15, 6.65
Slightly pleasant	11.55 (67)	10.08, 13.03
A little pleasant	13.15 (65)	11.31, 15.00
Mildly pleasant	21.73 (67)	19.43, 24.03
Moderately pleasant	36.39 (70)	33.43, 39.34
Very pleasant	63.59 (70)	60.39, 66.83
Strongly pleasant	66.94 (71)	63.89, 69.99
Extremely pleasant	81.55 (67)	79.18, 83.98
Most pleasant imaginable	97.24 (67)	95.04, 99.44

Note: The number of ratings used to compute the means and confidence intervals for each descriptor are in parentheses.

(a) *strongly unpleasant* vs. *very unpleasant*, $t(64) = 1.11$, $p > .27$, $d = 0.14$; (b) *a little unpleasant* vs. *slightly unpleasant*, $t(61) = -0.64$, $p > .52$, $d = 0.08$; (c) *a little pleasant* vs. *slightly pleasant*, $t(59) = 1.06$, $p > .29$, $d = 0.13$; and (d) *strongly pleasant* vs. *very pleasant*, $t(67) = 1.58$, $p > .11$, $d = 0.19$ (all significance tests were two-tailed). Also, the mean difference between *a little pleasant* and *mildly pleasant* yielded only a medium effect size, $d = 0.78$. These findings suggest that participants were unable to make distinctions between descriptors in these five pairs. Consequently, the descriptors *very unpleasant*, *a little unpleasant*, *a little pleasant*, and *very pleasant* were excluded from inclusion in the scale because the initial data screening revealed that more errors tended to occur when participants rated these descriptors than when participants rated their corresponding neighbours.

Observation of the remaining set of mean descriptor ratings suggested that their distribution about the *neutral* point was symmetric and that the pleasant and unpleasant version of each descriptor lay an equal distance from the *neutral* point. For this reason, the absolute values of the pleasant and unpleasant versions of each mean descriptor rating were averaged. In cases where the pleasant and unpleasant means consisted of an unequal number of ratings, the average of the two means was computed after weighting each accordingly. These averaged descriptor values were then linearly transformed to span a 100-unit continuum. After transformation, the descriptor values were rounded to the nearest whole number. The corresponding unit distance from the *neutral* 0-point (in both the unpleasant and pleasant directions) for each descriptor was as follows: *Barely* (7), *slightly* (12), *mildly* (24), *moderately* (38), *strongly* (70), *extremely* (85), and *most imaginable* (100). Finally, these descriptors were placed along the numeric scale continuum after removing the numeric markers. The resulting Empirical Valence Scale (EVS) is depicted in Figure 1.

DISCUSSION

Inspection of the EVS reveals that it shows similar characteristics to Green et al.'s (1993) LMS. On each scale, verbal descriptors appear to have a quasi-loglinear distribution along the scale continuum. For the EVS this distribution is particularly evident when comparing the distance between

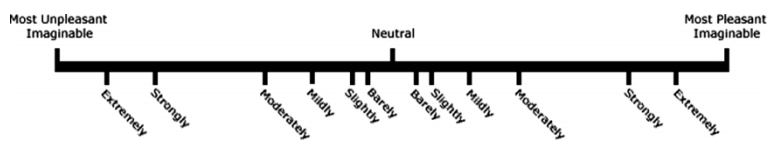

Figure 1. The Empirical Valence Scale (EVS).

no valence intensity (i.e., *neutral*) and *moderate* valence intensity (distance = 38 units) and the distance between *moderate* valence intensity and the *most imaginable* valence intensity (distance = 62 units). This inequality suggests that respondents perceive greater range in higher experiential valence than in lower experiential valence, which is obscured in LS where *moderate* valence intensity is typically designated as the midpoint between the *neutral* and anchor descriptors. These data do not support the assumption of equidistance in LS that assess emotional valence. It is possible that another set of descriptors might better approximate equidistance, but it is clear that the descriptors used in the vast majority of LS fail to do so.

These data also suggest that respondents make a distinction between an *extreme* and *most imaginable* intensity. Because VAS and LS of emotional valence typically use *extremely* or *very strongly* as anchors, ratings produced by these scales are likely susceptible to ceiling effects, a weakness that the EVS should minimise. Finally, like the LMS, the EVS employs a greater number of descriptors than typical VAS and LS, a characteristic that should help increase qualitative discrimination between different intensities of experienced valence.

Although a number of similarities between Green et al.'s (1993) LMS and the EVS exist, there are also some differences. First, the EVS contains an additional descriptor above the lowest intensity anchor point than does the LMS, and some of the descriptors are different. This difference results from a selection of descriptors that better capture characteristics of valence intensity in the present research. Also, participants in the present research were asked to rate almost twice as many descriptors. Doing so ensured that an adequate number of labels were available to maximise discrimination among qualitative intensities.

Second, and more importantly, the EVS and LMS differ in the relative spacing of descriptors. Specifically, a smaller portion of the EVS continuum reflects intensities above the *moderate* point than is seen with the LMS. Whereas the *moderate* point on the EVS is 38 units away from the 0-point (*neutral*), the *moderate* point on the LMS is only around 15 to 18 units away from the 0-point (based on visual inspection of Figure 1 in Green et al., 1993). This discrepancy may reflect differences between the scaling of descriptors in the hedonic and oral sensory modalities. However, it is also possible that slight differences in scale construction methodology may be responsible. Differences in the specific descriptors and number of descriptors used may have resulted in scale-specific framing effects. It also is notable that the EVS was developed with a bipolar format, whereas the LMS was developed with a unipolar format. Pilot data from our lab suggest that unipolar scale construction actually expands the relative distance between neutral and moderate ratings, and therefore cannot explain the greater distance between the *neutral* and *moderate* descriptors in the EVS relative to

the LMS. Finally, participants in the present research were college undergraduates and non-students recruited from the local community, whereas participants in Green et al.'s study were individuals who worked at a research institute, half of whom had prior experience with magnitude estimation of the rated modality.

The discrepancy in descriptor spacing between the EVS and LMS raises questions about the wisdom of applying the LMS, a scale developed for rating the intensity of oral sensation, to the hedonic sphere. As suggested by Bartoshuk et al. (2005), the LMS may outperform LS; however, if one wishes to utilise a scale with empirically spaced descriptors, it makes sense to use one in which the spacing is specifically tailored to the rating modality in question.

A concern may be raised that both the EVS and the LMS have a large gap between ratings of *moderate* and *strong*. Ideally, it would be useful to identify a descriptor that bisects these two points on the continuum. However, the English language may lack an appropriate descriptor that corresponds to this part of the continuum. A hybrid descriptor like *moderately strong* might be useful in this regard. A final limitation of the present research is that the normative sample consisted of only 74 participants, all of whom resided or went to school in one city. Use of a larger, more representative derivation group might produce a scale that more accurately reflects the true normative spacing of descriptors. However, increasing the sample size would likely lead to only minimal improvement in descriptor spacing, and would provide little practical benefit. Consequently, before worrying about more precise estimation of descriptors, it is important to determine whether the EVS does indeed provide a measurement advantage over existing traditional valence rating scales.

Implications and future directions

The EVS should offer a number of measurement strengths over non-empirically derived rating scales. First, the empirically derived distance between descriptor points should more accurately reflect normative perceived psychological distances between descriptor points. If so, compared to ratings made using traditional LS and VAS, EVS ratings should more closely approximate an interval-like scale of measurement. This property allows for stronger inferences regarding the quantitative magnitude of differences that exists between ratings. Second, the presence of multiple verbal descriptors should reduce cross-respondent differences in scale use, providing a distinct advantage over traditional VAS in which there are no intermediate labels. Third, a scale that includes anchors of maximal intensity should assess particularly intense experiences more accurately than traditional LS and VAS. Finally, empirically derived scales, such as the EVS, should enable one

to acquire both quantitative and qualitative information without compromising one in favour of the other.

Although the theoretical advantages of the EVS are clear, future research will need to evaluate its proposed measurement advantages. Because no gold standard exists for how one should obtain such validation, particularly with regard to evaluating the claim of interval-like scaling, it is important to consider some approaches that may provide insight into this issue. For example, following the presentation of emotional stimuli, one could evaluate whether use of the EVS results in a higher number of significant empirical associations between self-report ratings of valence and other indices of emotional responding relative to the use of traditional rating scales. For instance, one could determine if ratings made using the EVS are more likely to generate significant empirical associations with emotional trait ratings and physiological measures, which would be expected if the EVS does indeed provide enhanced quantitative measurement properties.

One also could examine whether use of the EVS leads to enhanced rater consistency in response to a common emotional stimulus relative to the use of traditional rating scales. One would expect enhanced interrater consistency if the EVS does indeed reduce differences in how respondents interpret the scale continuum. In addition, one could determine if ratings made using the EVS (relative to ratings made using traditional scales) are less susceptive to ceiling effects after exposure to stimuli that induce highly intense emotional experiences.

Regardless of how one wishes to examine the potential advantages of the EVS, it would be important to utilise traditional scales as comparisons. To draw the strongest conclusions from such comparisons it would be most ideal if the traditional scales were constructed in two different manners. First, comparison scales should be constructed in a manner that is typical to that used in previous research. Generally speaking, traditional LS and VAS lack anchor points that convey maximal intensity, and in the case of LS, consist of 3 to 5 intermediary verbal descriptors. If the EVS shows enhanced performance relative to these traditional scales, then additional comparisons should be made using traditional scales that consist of anchor points that convey maximal intensity, and in the case of LS, consist of identical intermediary verbal descriptors. Such an approach would provide insight into whether the EVS does indeed convey a measurement advantage and, if so, why.

Conclusions

The elusiveness of psychological constructs and processes requires continued improvement in measurement techniques. Consequently, the development and use of empirically derived rating scales may benefit affective science

specifically, and the entire field of psychology more broadly, because such scales may provide more sensitive quantitative and qualitative information than traditional rating scales. Hopefully, the present research will help pave the way for future improvements in the assessment of psychological experience.

Manuscript received 5 September 2006
Revised manuscript received 20 February 2007
Manuscript accepted 27 February 2007
First published online 31 July 2007

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*, 122–125.

Bartoshuk, L. M., & Marks, L. E. (1986). Ratio scaling. In H. L. Meiselman & R. S. Rivlin (Eds.), *Clinical measurement of taste and smell* (pp. 50–65). New York: Macmillan.

Benning, S. D., Patrick, C. J., & Lang, A. R. (2004). Emotional modulation of the post auricular-reflex. *Psychophysiology*, *41*, 426–432.

Borg, G. (1982). A category scale with ratio properties for intermodal and interindividual comparisons. In H. G. Geissler & P. Petxold (Eds.), *Psychophysical judgment and the process of perception* (pp. 25–34). Berlin, Germany: VEB Deutscher Verlag der Wisscnschaften.

Dimberg, U. (1997). Facial reactions: Rapidly evoked emotional responses. *Journal of Psychophysiology*, *11*, 115–123.

Green, B. G., Dalton, P., Cowart, B., Shaffer, G., Rankin, K., & Higgins, J. (1996). Evaluating the "Labeled Magnitude Scale" for measuring sensations of taste and smell. *Chemical Senses*, *21*, 323–334.

Green, B. G., Shaffer, G. S., & Gilmore, M. G. (1993). Derivation and evaluation of a semantic scale of oral sensation magnitude with apparent ratio properties. *Chemical Senses*, *18*, 683–702.

Hayes, M. H. S., & Patterson, D. G. (1921). Experimental development of the graphic rating method. *Psychological Bulletin*, *18*, 98–99.

Hu, S., & Wan, H. (2003). Imagined events with specific emotional valence produce specific patterns of facial EMG activity. *Perceptual and Motor Skills*, *97*, 1091–1099.

Izard, C. E. (1977). *Human emotions*. New York: Plenum Press.

Lang, P. J. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist*, *50*, 372–385.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological Review*, *97*, 377–395.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1992). A motivational analysis of emotion: Reflex-cortex connections. *Psychological Science*, *3*, 44–49.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1995a). *The International Affective Picture Systems (IAPS): Photographic slides*. Gainesville, FL: Center for Research in Psychophysiology, University of Florida.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1995b). *The International Affective Picture Systems (IAPS): Technical manual and affective ratings*. Gainesville, FL: Center for Research in Psychophysiology, University of Florida.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *22*, 1–54.

Marks, L. E., Borg, G., & Westerlund, J. (1992). Differences in taste perception assessed by magnitude matching and by category-ratio scaling. *Chemical Senses*, *17*, 493–506.

Marks, L. E., Stevens, J. C., & Bartoshuk, L. M. (1988). Magnitude matching: The measurement of taste and smell. *Chemical Senses*, *13*, 63–87.

Poultan, E. C. (1968). The new psychophysics: Six models for magnitude estimation. *Psychological Bulletin*, *69*, 1–19.

Schwerdtfeger, A. (2003). Using affective pictures instead of white noise: Still different response patterns for Petrie-style augmenters and reducers. *Personality and Individual Differences*, *34*, 253–262.

Snyder, D. J., Duffy, V. B., Hoffman, H. J., Ko, C.-W., & Bartoshuk, L. M. (2003). Age-related trends in food preference and intake. *Chemical Senses*, *28*, E55.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*, 153–181.

Stevens, S. S. (1971). Issues in psychophysical measurement. *Psychological Review*, *78*, 426–450.

Vanman, E. J., Dawson, M. E., & Brennan, P. A. (1998). Affective reaction in the blink of an eye: Individual differences in subjective experience and physiological reactions to emotional stimuli. *Personality and Social Psychology Bulletin*, *24*, 994–1005.