Teorie & Modelli, n.s., XII, 1-2, 2007 (269-276)

Test of the effect of scale labels on response linearity

Frederik Van Acker, Peter Theuns, Joeri Hofmans, & Olivier Mairesse[*] (Brussels)

Frisbie and Brandenburg (1979) and Dunham and Davison (1991) observed that many researchers spend a lot of time constructing questionnaire items that ensure maximal content and face validity while they overlook the fact that the labeling of the rating scale could cause invalid responses. The central issue of the present study is whether labels should be attached to all points of a scale or only to the extremes.

Frisbie and Brandenburg (1979) found small but significantly different means obtained using fully labeled scales and endpoints-only-labeled scales. When comparing two different formats of Likert-type scales Dixon, Bobo, and Stevick (1984) found that means obtained using these two kinds of labeling did not differ significantly. They found greater variability in ratings obtained with endpoints-only-labeled scales. Weng (2004) found that test-retest reliability is higher for fully labeled scales than for scales where not all response categories are labeled. These results show that labeling affects the mean and variability of responses. However, does labeling also affect the linearity of responses?

Several studies have pointed out that when subjects are asked to judge combinations of two or more stimuli, they use simple algebraic rules to integrate the subjective values of stimuli (Anderson 1981, 1982, 1996). These rules are revealed by the pattern of curves obtained from a factorial experimental design. Specific patterns of factorial curves occur when the rating scale is linear. Thus, to assess whether a particular scale is linear one needs only to test whether a pattern of factorial curves found previously for a specific integration rule also occurs when labels are attached to all points of the scale or only to the extremes. In this study we used this method to assess the interval properties of two differently labeled rating scales.

As part of an experiment on non-response in web surveys, 5829 students from the Vrije Universiteit Brussel were contacted via e-mail to par-

[*] Department of Work, Organizational and Economic Psychology, Free University of Brussels (VUB).

ticipate in an experiment. They were told that the study involved impression formation and attitude formation. Only 360 students completed the entire experiment, thereby providing the data considered in the present study. Respondents' ages ranged from 17 to 54 years (mean age 21.5 years, *SD* = 4.5). Of these participants, 129 indicated they had finished some form of higher education. In a between subjects design, participants were randomly assigned to one of two conditions that differed in the format of the rating scale: 180 participants received a fully labeled 7-point rating scale, the remaining 180 participants received a 7-point rating scale with only the endpoints labeled. In each condition the rating scale was presented vertically, below the stimuli. On average, respondents needed 15 minutes to complete the experiments. Total response times, when excluding the outliers, did not differ statistically depending on the type of labeling that was used.

Experiment 1

Zhu and Anderson (1991) and Dalal and Singh (1986) found that participants who rated overall job satisfaction, surmised from the integration of different aspects of some imaginary job, used an averaging rule with equal weights to integrate the different aspects of the job. In the present study we replicated these experiments. Since the averaging rule was well confirmed we expected to find parallel factorial curves provided that the rating scale was linear. Zhu and Anderson (1991) used a 20 cm line with numbers ranging from 0 to 20 and labeled only the endpoints, one with "extremely low" and the other with "extremely high". The 7-point rating scales used in the present study had a different number of labels depending on the condition a subject was assigned to. Finding an averaging rule depended on the capacity of these scales to elicit linear responses.

Procedure

Participants were presented a series of possible job situations and were asked how satisfied they would be in each particular situation. The stimuli comprised two dimensions: monthly salary and manager's leadership behavior, which were presented according to a factorial design. The levels of monthly salary were 0, 1000, 1500, 2000, and 2500 Euros and those of manager's leadership behavior were (1) "Your manager is authoritarian and dominant. The manager is not a group member but has a higher hierarchical position. Agreements are imposed by the manager and tasks are assigned by him/her", (2) "Your manager remains rather passive and lets things take

their course", and (3) "Your manager is someone who discusses things and makes agreements with you. Your manager is a group member and takes the

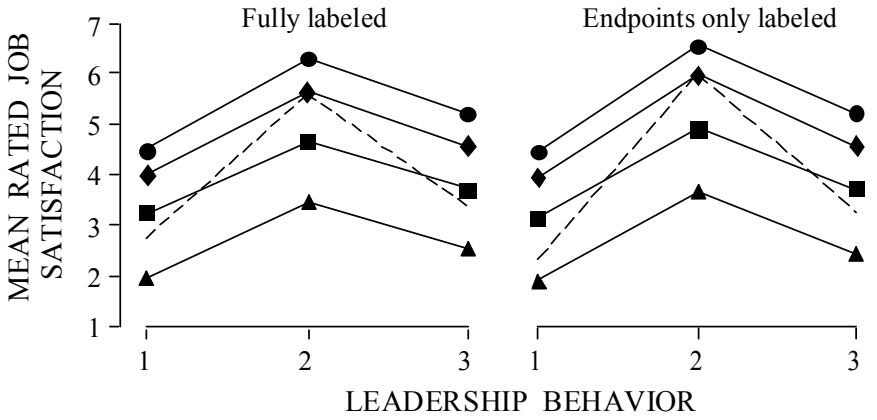

Figure 1. Results of Experiment 1. Mean rated job satisfaction from fully labeled (left) and endpoints-only-labeled rating scales (right) plotted against leadership behavior (1, 2, or 3) for different incomes [0 (dashed line), 1,000 (▲), 1,500 (■), 2,000 (◆), or 2,500 (●) Euros].

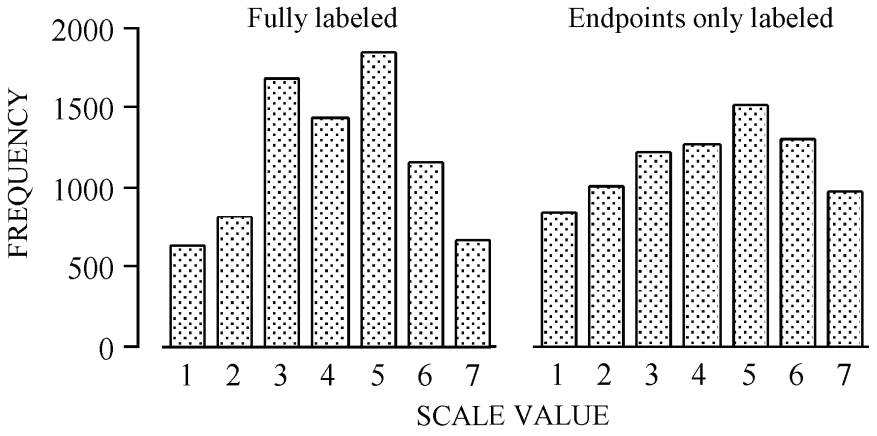

Figure 2. Results of Experiment 1. Distribution of ratings obtained with a fully labeled rating scale (left) and with a scale with only the endpoints labeled (right).

group members into consideration when making a decision". The series of 15 stimuli was presented three times with stimuli in random order. This resulted in 45 job satisfaction ratings for each participant.

Participants were assigned randomly to one of two conditions differing for the labeling of a 7-point rating scale. In one condition all scale points were labeled ("extremely dissatisfied", "very dissatisfied", "dissatisfied", "neutral", "satisfied", "very satisfied", "extremely satisfied") and in the other only the endpoints were labeled ("extremely dissatisfied" and "extremely satisfied").

Results and discussion

Figure 1 shows the results. As can be seen in the left diagram, with the exception of the curve for level 0 of monthly salary (dashed line), the data show a pattern of essentially parallel curves. Although visual parallelism is almost perfect, an analysis of variance with correction for non-sphericity showed a significant interaction [$F(5.1, 1074) = 7.2$, $p < 0.001$]. The right diagram for the fully labeled rating scale shows essentially the same pattern. Here too, although visual parallelism is almost perfect, there was a significant interaction [$F(5.1, 1074) = 6.9$, $p < 0.001$]. Considering the large number of participants, these interactions are most probably due to some small context effect.

We had hypothesized that the data would fit an averaging model with equal weights. We thus expected the curve for the level 0 of monthly salary to cross over other curves. As can be seen in Figure 1, the crossover occurred in agreement with earlier work by Zhu and Anderson (1991) and Dalal and Singh (1986). We may thus conclude that 7-point rating scales yield essentially linear responses both when all their points are labeled and when only their endpoints are labeled.

Figure 2 provides an overview of the distributions of responses across the two rating scales. In agreement with the result of Dixon, Bobo, and Stevick (1984), the distribution of responses obtained from the endpoints-only-labeled scale was significantly flatter ($z = 3.8$, $p < 0.001$, Kolmogorov-Smirnov test). That is, participants tended to avoid endpoints more when a fully labeled scale was used.

Experiment 2

Participants were asked to judge the attractiveness of people from pictures of a person's face combined with a personality trait. Anderson (1965) had men and women judge persons based on a set of four personality traits.

He found that the integration of these traits followed an averaging rule with equal weights. We then expected that participants confirmed this finding in the case their response function, and thus the rating scale, was linear.

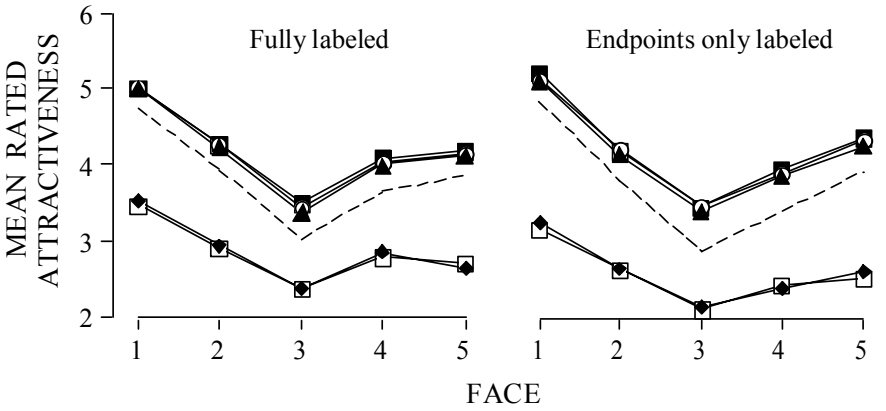

Figure 3. Results of Experiment 2. Mean rated attractiveness from fully labeled (left) and endpoints-only-labeled rating scale (right) plotted against stimulus face (1 to 5) for each associated personality trait [no trait (dashed line), unfriendly (◆), friendly (■), dishonest (□), honest (○), intelligent (▲)].

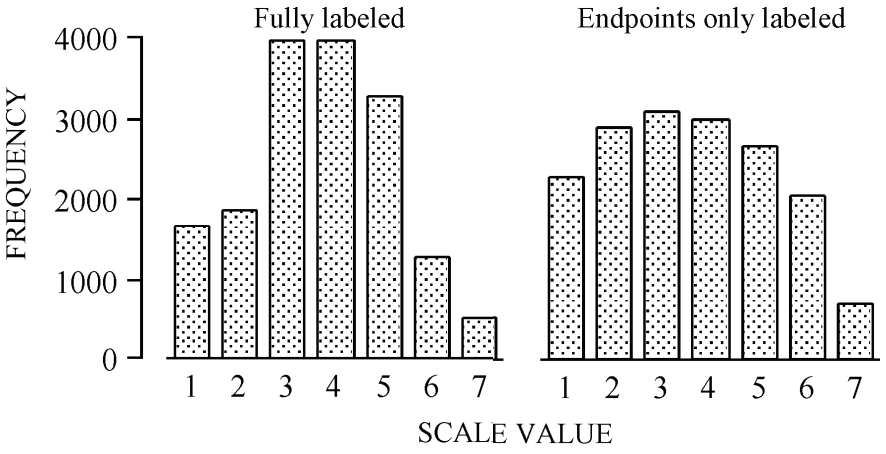

Figure 4. Results of Experiment 2. Distribution of ratings obtained with a fully labeled rating scale (left) and with a scale with only the endpoints labeled (right).

Procedure

The experiment was set up as a 5 (face) × 6 (trait) × 3 (repetition) design. We used five faces covering a somewhat large range of attractiveness, formerly used by Braun, Gruendl, and Marberger (2001). The personality trait was: intelligent, honest, dishonest, friendly, or unfriendly. Each face was also presented without any personality trait. The series of these 30 combinations was presented three times with combinations in random order. The participants that were in the "endpoints-only-labeled" condition of Experiment 1 now used the 7-point rating scale with only the endpoints labeled, one "extremely unattractive" and one "extremely attractive". The participants that were in the "fully labeled" condition of Experiment 1 now used the 7-point rating scale with its points labeled respectively "extremely unattractive", "very unattractive", "rather unattractive", "neutral", "rather attractive", "very attractive", and "extremely attractive".

Results and discussion

Figure 3 shows the results. The left and right diagrams show that, essentially, the stimuli were integrated additively. The analysis of variance does not support parallelism both when only the scale endpoints were labeled and when all scale points were labeled [$F(6.9, 2864) = 14.4$ and $F(7.1, 2864) = 10.7$, $p < 0.001$, respectively]. As observed before, these interactions are most probably due to some small context effect. These results lead us to conclude that both kinds of labeling elicit essentially linear responses.

The curve for the "no trait" level (dashed curve) is expected to cross the other curves if the integration rule is a weighted average. In Figure 3, visual inspection shows that this prediction is not confirmed. This is in contrast with the findings of Lampel and Anderson (1968) who found an averaging rule for the integration of visual information and personality-trait adjectives. The present results agree more with the possibility that the integration rule is an addition rather than an average.

Figure 4 shows the frequency distributions of responses in the two labeling conditions. In agreement with the results of Experiment 1, the distribution of responses from the endpoints-only-labeled scale was significantly flatter ($z = 9$, $p < 0.001$, Kolmogorov-Smirnov test).

General discussion

To assess whether a 7-point rating scale yielded linear responses, we tested whether a pattern of parallel factorial curves found in previous litera-

ture also occurred when labels were attached to all points or only to the extreme points of the scale. We found that both of these scales yielded a pattern of essentially parallel factorial curves, indicating that the scales were essentially linear. Although this probably has no implications for the validity of the current study, we were not able to replicate the expected integration rule in the second experiment. One possible explanation for this finding is that data collected by means of the internet are more biased due to context effects. The statistically significant interaction in both experiments also points in this direction. Moreover, only about 6 percent of the contacted sample yielded data that were suitable for analysis. Therefore we suggest that future research assesses whether results obtained from functional measurement experiments conducted on the web can yield valid results.

References

Anderson, N. H. (1965). Averaging versus adding as a stimulus-combination rule in impression formation. *Journal of Experimental Psychology, 70*, 394-400.

Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.

Anderson, N. H. (1982). *Methods of information integration theory*. New York: Academic Press.

Anderson, N. H. (1996). *A functional theory of cognition*. Mahwah, NJ: Erlbaum.

Braun, C., Gruendl, M., Marberger, C., & Scherber, C. (2001). Beautycheck - Ursachen und Folgen von Attraktivitaet. Report. [pdf-document]. Retrieved September 26, 2006 from: http://www.beautycheck.de/english/bericht/bericht.htm

Dalal, A. K., & Singh, R. (1986). An integration theoretical analysis of expected job attractiveness and satisfaction. *International Journal of Psychology, 21*, 555-564.

Dixon, P. N., Bobo, M., & Stevick, R. A. (1984). Response differences and preferences for all-category-defined and end-defined Likert formats. *Educational and Psychological Measurement, 44*, 61-66.

Dunham, T. C., & Davison, M. L. (1991). Effects of scale anchors on student ratings of instructors. *Applied Measurement in Education, 4*, 23-35.

Lampel, A. K., & Anderson, N. H. (1968). Combining visual and verbal information in an impression-formation task. *Journal of Personality and Social Psychology, 9*, 1-6.

Frisbie, D. A., & Brandenburg, D. C. (1979). Equivalence of questionnaire items with varying response formats. *Journal of Educational Measurement, 16,* 43-48.

Weng, L. (2004). Impact of the number of response options of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64*, 956-972.

Zhu, S., & Anderson, N. H. (1991). Self-estimation of weight parameters in multiattribute analysis. *Organisational Behavior and Human Decision Processes, 48*, 36-54.

Abstract

This paper reports two experiments designed to test whether adding labels to a 7-point rating scale affects the linearity of this scale. We tested whether a pattern of factorial curves found in previous studies for a specific integration rule also occurred when labels were attached to all points of the scale or only to the extremes. The results show that fully labeled rating scales and scales that have labels only for the extreme points reproduce the pattern of factorial curves found in previous studies. We conclude that both of these scales yield linear responses.

Riassunto

Questo articolo riporta due esperimenti effettuati per controllare se l'aggiunta di etichette ad una scala di valutazione a 7 punti influenza la linearità di tale scala. Abbiamo controllato se una configurazione di curve fattoriali per una regola di integrazione specifica, che era stata trovata in studi precedenti, si verifica sia con etichette attaccate a tutti i punti della scala che con etichette attaccate solo agli estremi. I risultati mostrano che sia le scale complete di etichette che le scale etichettate solo agli estremi riproducono la configurazione di curve fattoriali trovata in studi precedenti. Si conclude che entrambi tali scale producono risposte lineari.

Aknowledgment. I would like to thank Norman H. Anderson for his valuable comments on an earlier version of this paper.

Addresses. Frederik Van Acker (frederik.van.acker@vub.ac.be), Peter Theuns (peter.theuns@vub.ac.be), Joeri Hofmans (joeri.hofmans@vub.ac.be), Olivier Mairesse (olivier.mairesse @vub.ac.be): Department of Work, Organizational and Economic Psychology, Free University of Brussels (VUB), Pleinlaan 2, B-1050 Brussel, Belgium