Teorie & Modelli, n.s., XII, 1-2, 2007 (259-268)

An alternative for prescribed integration rules in testing the linearity of a response measure

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

How can one test linearity of a given response measure? This question has received attention with the increased study of algebraic models. The advantage of studying algebraic models is that they permit simultaneous analysis of the algebraic model and of the responses generated by the model (Anderson, 1981, 1982, 1992, 1996; Birnbaum & Veit, 1974; Weiss, 1972). This means that if the data fit a hypothesized model, this fit is perceived as joint support for both the model and the linearity of the response scale. In experiments studying algebraic models, multiple factors are manipulated simultaneously. The stimuli manipulated by the experimenter (φ_i) are transformed into subjective stimuli (s_i) by the Valuation-function. Then these subjective stimuli result in a subjective response (r), based upon a theory of integration, of which previous research has shown that it can be described in terms of a simple algebraic model. Finally, the subjective response becomes an overt response through the Response-function (R), for example a rating on a category rating scale.

Three simple algebraic models are frequently found to be good descriptors of integration rules for most judgmental tasks: the adding-type model, the multiplicative model, and the averaging model (Anderson, 1981, 1982, 1992, 1996; Weiss, 2006). These three algebraic models predict specific patterns in the data if the Response-function (R) is linear, i.e., if the rating is made on a linear scale. For the adding-type model and the averaging model with equal weights, one should observe a pattern of parallelism when plotting the data in a factorial graph whereas the multiplicative model predicts a linear fan pattern (Anderson, 1981, 1982, 1992, 1996; Graesser & Anderson, 1974; Weiss, 2006; Weiss & Shanteau, 1982).

Although the logic seems very convincing at the same time there seems to be some circularity in the reasoning (Birnbaum, 1982, p. 452; Weiss, 2006, p. 213). A test on the integration rule assumes a linear response scale

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

whereas testing the linearity of the response scale places some assumptions on the integration rule. Because the integration rule and the response scale are validated simultaneously, one has to deal with what is called the "problem of evidence" (Anderson, 1996). This problem points to the fact that a specific data pattern offers strong support for the matching integration rule but is no absolute proof. For example, observed parallelism provides strong support for an adding-type integration rule and for linearity of the response measure, but it cannot prove both. In fact, all possible combinations of a certain integration rule and a response measure can be summarized in the four cases shown in Table 1.

Case	Adding-type integration rule	Linear response measure	Observed pattern
1	Yes	Yes	Parallelism
2	Yes	No	Nonparallelism
3	No	Yes	Nonparallelism
4	No	No	Parallelism or nonparallelism

Table 1. The status of the integration rule and response measure related to the pattern observed in the factorial graph.

The first case is the preferred one: the integration rule is indeed an adding-type model and the response measure is linear. In this case we are guaranteed to find a pattern of parallelism in the factorial plot, supporting both premises jointly and hence each of them separately (Anderson, 1996). The non-parallelism resulting from Cases 2 and 3 is obvious but does not allow one to discriminate between Cases 2 and 3 without further experimental manipulations. Even Case 4 is likely to generate non-parallelism although parallelism exceptionally can occur in case a nonlinear response scale per chance compensates for a non-additive integration rule thereby yielding parallelism in the factorial graph (Anderson, 1996). In summary, parallelism can be obtained in Cases 1 and 4, while non-parallelism can be found with Cases 2, 3 and 4. The major problem is then, given the pattern found in the raw data, to be confident in the cause of this specific pattern. In other words, when we find a pattern of parallelism, it needs to be ensured that this is due to Case

1, and not to Case 4. On the other hand, when we perceive non-parallelism, we need to determine whether this is due to the non-additive integration rule, to a nonlinear response measure, or both.

In one approach to resolve the "problem of evidence", respondents are presented with an information integration task, i.e., they are asked to give the average, the difference, or the ratio of two stimuli (Weiss, 1972; Anderson, 1996, p. 95). The logic behind this approach is that when data from one response modality fit the imposed integration rule while data from the other modality do not, the latter is charged with invalidity or non-linearity (Weiss, 1972). Although the rationale is quite sensible, this methodology has an important drawback, i.e., respondents could use an integration rule different from the one imposed by the experimenter. In a series of experiments, Birnbaum and Veit (1974) and Veit (1978) manipulated the nature of the information integration task by having the respondents lift weights simultaneously in both hands and judge either the difference, the ratio, or the average heaviness of both weights. Based on the principle of scale convergence, or the independency of the stimulus values on the integration task, the scale values derived with each integration task ought to agree (Anderson, 1972). In their papers Birnbaum & Veit (1974) and Veit (1978) reported two major findings. First of all, both the difference model and the ratio model were supported by the raw data, showing a pattern of parallelism and a linear fan respectively. Secondly, the scale values agreed only if the data from the subtractive model and the data from the ratio model were both fitted to the same subtractive model. Based on both findings they concluded that respondents apply the same integration rule, i.e., a subtractive model, regardless whether they are instructed to rate differences or ratios (Birnbaum & Veit, 1974; Veit, 1978). Moreover, Birnbaum and Veit (1974) and Veit (1978) concluded that the use of a single factorial design does not suffice to resolve the "problem of evidence".

Although the principle of scale convergence hands an additional criterion to attack the "problem of evidence" it is very time-consuming to test such a series of experiments. Moreover the criterion of scale convergence is useless in situations where the parameters refer to molar units depending on the context, as in decision theory and social judgment (Anderson, 1982, p. 200). In these situations, the functional value of any stimulus depends on the goal, and will not be constant across different tasks.

The approach adopted in this paper tries to avoid the problems caused by making inferences based on a single information integration task, i.e., the need to rely on the assumption that some imposed integration rule is adopted by the respondents. Furthermore it is far less time-consuming than conducting a series of experiments and applying the criterion of scale convergence. In the experiments described in this paper we presented the participants with specific physics problems. For example, in Experiment 1 the participants rated the weight of a beam while the material and the height of the beam were manipulated. After the experiment, each participant was interviewed about their knowledge of the problem. Depending on whether the participant knew the formula for weight and, by inference, the appropriate integration rule we expected a certain pattern in the raw data. For example if a participant knew that volume × density equals mass, we expected a linear fan in his/her raw responses, if the participant thought the formula was volume + density then a pattern of parallelism should have appeared. If the respondents use the same integration rule consistently, then only a nonlinear scale could cause deviations from these predicted patterns.

Experiment 1

Method

A group of 32 undergraduate physics students with mean age of 21.4 years, and a standard deviation of 2.7 years, took part in Experiment 1. All participants were paid 10 Euros in return for their participation. We asked the participants to rate the weight of a beam displayed on a computer monitor in front of them. The beams to be rated were constructed according to a 4 × 4 factorial design and all stimuli were presented three times and in random order. The depth and width of the beam were fixed at 10 cm. One factor was the height of the beam with levels being 10, 20, 30, and 40 cm and the other was the material of the beam with levels being lead, clay, ice, and Styrofoam. Before the experiment started the participants were presented with the four materials and the four different heights. The participants were then asked to indicate the lightest and the heaviest beam based on the information provided. Afterwards, these descriptions were placed as end anchors on their graphical rating scale, a 625 × 17 pixel slider on a 1024 × 768 pixel monitor. After completing the experiment, each participant was asked to explain how to compute the mass of a beam.

Results and discussion

Based on the information provided in the exit-interview, 27 participants reproduced the correct formula to compute the mass of a beam, i.e., mass = volume × density. Consequently, we expected a linear fan pattern in the raw data of these participants. This linear fan is clearly demonstrated in the left diagram in Figure 1 which is a factorial plot of the data of these 27

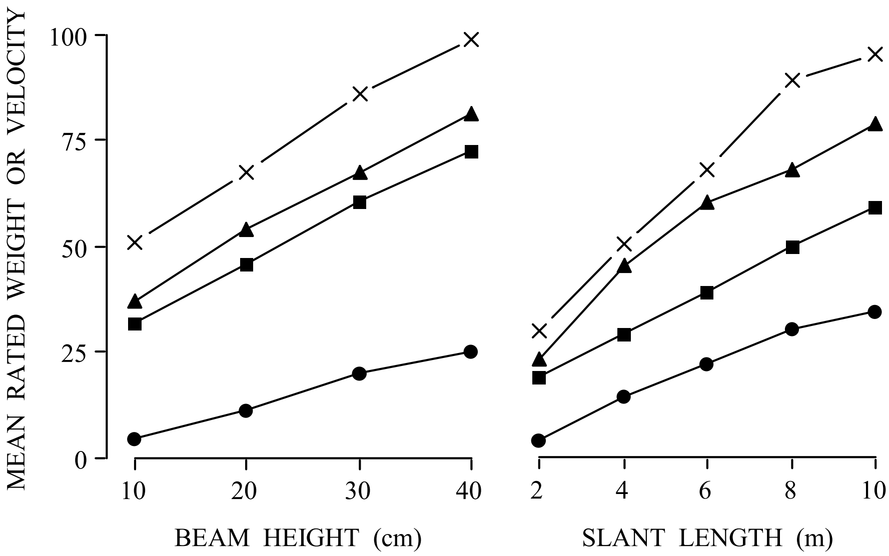

Figure 1. Left: results of Experiment 1 – mean rated weight of beam plotted against beam height for different beam materials [lead (×), clay (▲), ice (■), or Styrofoam (●)]. Righ: results of Experiment 2 – mean rated velocity of ball plotted against slant length for different ball masses [70 (×), 50 (▲), 30 (■), or 10 Kg (●)].

participants averaged over repetitions and participants. The multiplicative model was supported by a significant bilinear interaction [$F(1, 31) = 54$, $p < 0.001$].

The other five participants correctly stated that both height and density impact on the weight but failed to spontaneously give a formula linking both attributes. Probably these participants would have succeeded in providing a formula with a little support but the fact that it did not come spontaneously prevented us from making a specific hypothesis concerning their integration rule. Therefore, the responses of these five participants were not analyzed.

Observing a linear fan in the raw data has two implications. First of all it supports the multiplicative nature of the integration rule. In this experiment, multiplicativity was expected because these participants knew that volume and density must be multiplied to determine the mass of an object. Since the integration rule is under control, the patterns in the raw data only appear if the Response-function is linear since deviations from linearity would cause deviations from the predicted pattern. Thus, the results of this experiment validate the graphical rating scale as a linear response measure.

Experiment 2

Method

The same group of 32 undergraduate physics students participated in a second experiment where their task was to rate the velocity of a ball rolled from a slant. The mass of the ball and the length of the slant were manipulated according to a 4 × 5 factorial design. The levels of the mass of the ball were 10, 30, 50, and 70 kg and those of the length of the slant were 2, 4, 6, 8, and 10 m. The degree of inclination of the slant and the volume of the ball were fixed during the entire experiment. During the introduction prior to the experiment it was mentioned that both the length of the slant and the mass of the ball were manipulated and the levels of both factors were presented. All stimuli appeared three times and in random order. Ratings were made using the graphic rating scale that was used in the first experiment. The participants were first asked to indicate the slowest and the fastest ball. Subsequently these descriptions were used as end anchors on the participants' graphical rating scale. Analogue to Experiment 1 we asked in an exit-interview to explain how to compute the velocity of a ball rolling down a slant.

Results and discussion

Based on the information provided in the exit-interview, we discerned two different groups of participants. A first group consisted of nine participants who correctly attributed the velocity of the ball to the length of the slant thereby disregarding mass as an influential factor. An analysis of the raw data of these participants confirmed that they did not take the mass of the ball into consideration when rating the velocity since the main effect for mass was not significant [$F(3, 24) = 0.16$] while the effect of the length of the slant was significant [$F(4, 32) = 63, p < 0.001$].

Another group of seven participants thought that the velocity was a multiplicative function of the mass of the ball and the length of the slant. Since these participants expected both factors in the design to integrate in a multiplicative manner, we expected to see a linear fan pattern in the raw data. The right diagram in Figure 1 displays a factorial plot of the data averaged over repetitions and participants. In line with the predictions of the multiplicative model, the interaction was concentrated in the linear by linear component [$F(1, 6) = 15.6, p < 0.01$].

The remaining 16 participants were unable to provide a formula spontaneously. They mentioned something like “both weight and height matter” making it impossible to make predictions about the integration rule and the matching pattern in these participants’ data.

This experiment demonstrates that even incorrect formulas can provide useful information when the object of interest is the linearity of the response scale. The correct formula for velocity was provided spontaneously by some participants and this (non-)integration rule was confirmed in an analysis of their raw data. This finding is interesting in itself but not very helpful for our purposes since these results would be obtained even with a monotone scale. The results of the second group of seven participants are more important. Since this group consisted of participants who, based on the information provided in the exit-interview, thought that mass and length of the slant combined multiplicatively, a linear fan was hypothesized for this group. The observed linear fan pattern in the raw data provides joint support for a multiplicative integration of both attributes and a linear response scale. Since a linear fan would hardly be obtained unless the response measure was linear, confidence can be put in the linearity of the graphical rating scale used in this experiment.

General discussion

The method adopted in this paper proves to be a useful approach in resolving the “problem of evidence”. Where in “traditional” functional measurement experiments the integration rule and the response measure are validated simultaneously, this approach controls for the former. Therefore only a nonlinear response scale could cause deviations from the predicted pattern. A similar reasoning is used as with information integration tasks where participants are instructed to apply a certain integration rule. However, a major disadvantage with such tasks is the need to rely on the assumption that the participants indeed integrate the stimuli as instructed by the experimenter. Research of Birnbaum and Veit (1974) and Veit (1978) showed that a subtractive model is applied regardless of whether participants are instructed to rate ratios or differences. This means that the assumption of adoption of the imposed integration rule by the participants can not be surmised. Fortunately, functional measurement hands another criterion allowing a test of this assumption. If the algebraic model is correct and if the rating scale is linear, then the functional stimulus scales constitute an interval scale of the stimuli being measured. A consequence is that functional stimulus scales from different experiments should converge (Anderson, 1972;

Birnbaum & Veit, 1974; Veit, 1978). When functional stimulus scales obtained with two or more information integration tasks are linearly related, the integration rule in each experiment as well as the response scale is validated simultaneously. However, if the stimulus scales relate in a nonlinear way, additional tests are necessary to locate the source of this incongruence. Since applying this criterion implicates the implementation of several experiments this approach is very time-consuming.

The approach tested in this study tries to resolve for the disadvantages of experiments with prescribed integration rules, but along with its advantages this approach has a number of limitations as well. First of all it is limited to stimuli for which the integration rule can be discovered in one way or another, in practice this means that it is probably limited to physics problems. Second, the data obtained from some participants are useless, at least given our objectives, because these participants fail to give a formula. This does not mean that these people do not intuitively apply the correct integration rule (Karpp & Anderson, 1997) but, due to their failure to give a formula, we have no specific hypothesis regarding their integration rule. Furthermore, it is impossible to know in advance how many participants each group will contain since this depends on the formula they give in the interview. In the cases when very few people refer to a particular formula there may be problems regarding statistical power. Therefore, just as in all functional measurement experiments, it is advisable to work with strong designs and by preference single-subject designs. A final disadvantage is the necessary assumption that the respondents use the same integration rule consistently in both tasks. Research from Krist, Fieberg, and Wilkening (1993) and Wilkening and Martin (2004) has demonstrated a dissociation between judgements (verbal performance) and actions (motor performance) in intuitive physics. This means that one should be careful when using two tasks requiring totally different response measures since then dissociation may appear, violating the assumption of independence of the integration function from the response measure. However, this problem is of minor importance in our research since both responses were judgements, i.e., verbal performances, one requiring the formulation of a formula and one involving a response using a graphical rating scale.

Summarizing, our results show that the approach adopted in this paper can be useful in testing the linearity of various response instruments. In line with previous research on graphical rating scales (Anderson, 1982, p. 7; Weiss, 1972) the self-anchoring graphical ratings used in the present experiments proved to be linear response measures.

References

Anderson, N. H. (1972). Cross-task validation of functional measurement. *Perception & Psychophysics, 12*, 389-395.

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. (1992). Integration psychophysics and cognition. In D. Algom (Ed.), *Psychophysical approaches to cognition* (pp. 13-113). Amsterdam: Elsevier.

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

Birnbaum, M. H. (1974). The nonadditivity of personality impressions. *Journal of Experimental Psychology, 102,* 543-561.

Birnbaum, M. H. (1982). Controversies in psychological measurement. In B. Wegener (Ed.), *Social attributes and psychological measurement* (pp. 401-485). Hillsdale: Erlbaum.

Birnbaum, M. H., & Veit, C. T. (1974). Scale convergence as a criterion for rescaling: Information integration with difference, ratio, and averaging tasks. *Perception & Psychophysics, 15,* 7-15.

Graesser, C. C., & Anderson, N. H. (1974). Cognitive algebra of the equation: Gift size = generosity × income. *Journal of Experimental Psychology, 103*, 692-699.

Karpp, E. R., & Anderson, N. H. (1997). Cognitive assessment of function knowledge. *Journal of Research in Science Teaching, 34*, 359-376.

Krist, H., Fieberg, E. L., & Wilkening, F. (1993). Intuitive physics in action and judgment: The development of knowledge about projectile motion. *Journal of Experimental Psychology, 19*, 952-966.

Veit, C. T. (1978). Ratio and subtractive processes in psychophysical judgment. *Journal of Experimental Psychology: General, 107*, 81-107.

Weiss, D. J. (1972). Averaging: An empirical validity criterion for magnitude estimation. *Perception & Psychophysics, 12*, 385-388.

Weiss, D. J. (2006). *Analysis of variance and functional measurement: A practical guide*. New York: Oxford University Press.

Weiss, D. J., & Shanteau, J. C. (1982). Group-Individual POLYLIN. *Behavior Research Methods & Instrumentation, 14*, 430.

Wilkening, F., & Martin, C. (2004). How to speed up to be in time: Action-judgment dissociations in children and adults. *Swiss Journal of Psychology, 63*, 17-29.

Abstract

One method to test whether graphical ratings are linear response measures is to prescribe an integration rule and test whether the resulting pattern of factorial curves is that predicted by this rule. This method does not guarantee that respondents do not use an integration rule different from the prescribed rule. To resolve this problem, we had participants graphically rate either the weight of a beam varying in density and volume or the velocity of balls of different masses rolling down slants of various lengths. After the experiment, we asked participants whether they knew the formulas to calculate beam mass and ball velocity. The participants who knew the correct formulas produced patterns of factorial curves in agreement with the formulas. These results confirm that graphical ratings are linear response measures.

Riassunto

Un metodo per controllare se le valutazioni grafiche sono misure lineari è quello di prescrivere una regola di integrazione a controllare se le curve fattoriali hanno la configurazione prevista da tale regola. Questo metodo non garantisce che i partecipanti non usino una regola di integrazione differente da quella prescritta. Per risolvere questo problema, abbiamo chiesto a dei partecipanti di valutare graficamente sia il peso di una trave con densità e volume varianti che la velocità di una palla di massa variante che rotolava lungo un piano inclinato di lunghezza variante. Dopo l'esperimento, abbiamo chiesto ai partecipanti se conoscevano le formule per calcolare la massa della trave e la velocità della palla. I partecipanti che conoscevano le formule corrette produssero configurazioni di curve fattoriali in accordo con le formule. I risultati confermano che le valutazioni grafiche sono misure lineari.

Aknowledgments. Supported by Grant OZR1041BOF of Vrije Universiteit Brussel. We wish to thank Norman Anderson for helpful comments on an earlier version of the manuscript.

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