*Slawomir Pasikowski*
Poland

The New Educational Review

# The Problem of Matching Rating Scales in Educational Measurement of Variables Modelled as Sets of Oppositional Pairs

DOI: 10.15804/tner.2018.54.4.22

**Abstract**
The validity of educational measurement of characteristics modelled in the structure of oppositional pairs is determined, among other things, on matching the rating scale to the properties of the operationalised variables. 270 people participated in the study of this issue. The results revealed the significance of the type of rating scale and its length in determining the results of characteristics measurement.

**Keywords:** *rating scales, scales of measurement, validity, oppositions*

## Introduction

The validity of educational measurement depends mainly on a proper conceptualisation of the measured characteristics, and its operationalisation. Numerous features which are the subject of interest of studies on education have multidimensional backbone, e.g., knowledge, attitudes, self-assessment, well-being, self-image and the image of others, and satisfaction. In practice, many of them are measured using bipolar rating scales, which expresses the researcher's assumption about the dichotomous structure of the characteristics. The example of an adaptation of such scales and research assumptions is the semantic differential scale. However, the selection of a rating scale is not always suited to the scale of measurement

chosen by the researcher or the adopted multidimensional structure of the characteristic. A serious difficulty in the case of measurement, not only with respect to multidimensional characteristics, is surjection, i.e., an instance in the process of modelling characteristics in a system of symbols where the same symbol represents different intensity or states of the measured characteristic. The issue of surjection is strongly tied to the problem of the mid-point on a rating scale. This mid-point is often identified as the one that best corresponds to the intensity of a characteristic by those who indicate a similar intensity of both oppositional tendencies, and by those who declare a lack of both tendencies or their negligible intensity. Nevertheless, in research practice, the use of bipolar scales in measuring complex characteristics is still popular.

## Mapping functions, measurement scales and rating scales

The subject of educational measurement are usually latent variables, i.e., those which are not directly observable. Measurement of such characteristics requires prior construction of a model representing a given characteristic, and testing its fit to sampled empirical data. Measurement consists in matching symbols to specific values of a characteristic according to established rules, which is described in metrology and psychometry as a mapping function or simply as mapping (Chadha, 2009). In other words, measurement is a process in which mapping occurs from one set (domain) onto another set (codomain). The first set represents the values of a characteristic. The other is a system of symbols used to map this characteristic and its values. The fundamental expectation in the measurement process is the mapping of a characteristic in such a way that specific values of the characteristic respond to specific symbols. This expectation is connected with the term unequivocality of mapping. Different values of the characteristic should, therefore, be represented by various symbols. This makes the measurement accurate, and its results can be used to perform the operation of comparison. The precision of this comparison depends on the adopted scale of measurement. In the case of measurement tests, this usually means continuous features, for the mapping of which researchers usually use an interval scale of measurement, which assumes a fixed unit of measurement. This type of scale allows them to conduct linear transformations and to determine the distance between the values of a characteristic, e.g., between the result of measurement of person A and the result of measurement of person B. Scales of measurement with a lower level of precision, such as ordinal or nominal, do not have this feature.

Assuming a specific scale of measurement codetermines the selection of rating scales. In the case of measuring continual characteristics, the most often applied rating scales are discrete (numerical scales, such as a Likert-type scale) and continuous (Visual Analogue Scale, in its simplest form as a straight horizontal line). When the modelled characteristic is in an opposition structure, each type is used in one of two variants, i.e., bipolar or unipolar. In the first instance, the poles of the rating scale represent opposite aspects and tendencies. Thus, such a scale has two anchor points. In the second instance, each oppositional component is assessed using a separate rating scale, with a single anchor point – one pole signifies minimum intensity and the other maximum intensity of the tendency.

## The problem of matching a rating scale

Apart from matching the model to the measured phenomenon, one should factor in convergence of the adopted scale of measurement and the adopted rating scale. Otherwise, the validity of the measurement may decrease, as researchers sometimes lose sight of this convergence when deciding on the type of rating scale which provides data characteristic to the less precise scale of measurement than the one which was adopted at the starting point. An example of this is age, operationalised as a continuous characteristic and measured using a discrete rating scale, which allows the respondents to choose the range in which their age is located. A similar instance is conducting measurement of a continuous characteristic using a single item equipped with a discrete rating scale, which is not enough to reach a level that guarantees the possibility of implementing a fixed unit of measurement. What is more, discrete scales provide data from an ordinal scale of measurement, which results in limitations in modelling the characteristic, and in selecting the methods of statistical analysis. However, the assumption that a continuous characteristic can be modelled using a sum of values marked on discrete rating scales of a given test allows us to overcome these limitations. Hence, discrete scales are becoming very competitive against continuous scales. This stems mainly from their comfortable implementation, including their intuitive reading by respondents. However, it was noticed that continuous scales provide more varied results than discrete scales. That is why they are preferred in measurements in which reflecting individual differences is of crucial importance (DeVellis, 2017). Notwithstanding, there are studies which demonstrate that discrete scales provide more stable measurement results (Svensson, 2000), which is quite obvious given the average length of these scales – usually around 3–7 points – and the fact that

respondents can easily remember the marked number. All in all, opinions on the possibility of indicating a generalised superiority of discrete or continuous rating scales differ. What is highlighted is the significance of objectives and conditions in which specific types of rating scales are implemented (ibidem).

A separate issue is the length of the rating scale. It is not surprising that characteristics of discrete rating scales change along with an increase in their length. As a result, the term quasi-continuous scale was introduced (Hadijsky, 2007). There are also strong positions concerning the optimal length of rating scales (Bandalos, 2018; Preston & Colman, 2000). According to them, it oscillates between 5 and 11 points in the case of discrete scales and 100–150 points (mm) in the case of continuous scales (DeVellis, 2017).

In the measurement of characteristics modelled in the structure of oppositional pairs, the issue of anchoring the scale is significantly more important than its length. Double-anchored scales, i.e., bipolar scales, simultaneously refer to two aspects of a given characteristic. Such scales allow for observing the shaping of a complex characteristic to a limited degree. They only allow for assessing the dominating aspect and leave the issue of selecting the mid-point unresolved. Hence, in situations requiring adequate measurement precision it is preferable to implement unipolar scales. Each aspect, which corresponds to one pole in bipolar scales, is represented by an entire continuum of a rating scale. This type of solution requires the implementation of suitable data-integrating formulas. However, this is a separate issue which will not be discussed here.

## The issue of the mid-point on a scale as an empirical case of *tertium non datur*

A key issue in selecting the mid-values in a bipolar rating scale rests in the occurrence of surjection. This mid-point is often identified as one that best corresponds to the intensity of a characteristic by those who demonstrate similar intensity in both oppositional tendencies, and by those who declare a lack of both tendencies or their insignificant intensity.

The issue of the mid-point on a rating scale was noticed long ago, and effective solutions were developed in response (Priester & Petty, 1996) and this is not about the proposed ipsative scale. However, the common practice in research is to use bipolar scales to assess the occurrence and intensity of complex characteristics, despite the obvious risk of measurement inaccuracy in situations of similar intensity of separate tendencies within a given characteristic. The presented study is to

illustrate this risk and the role of selected qualities of the type of rating scale in determining measurement results. From among the possible characteristics, an attitude has been chosen for this purpose, as it is relatively commonly modelled in the structure of the opposition, exactly its sign: positive-negative. Moreover, it is relatively easy for the respondent to imagine and is accessible to common individual experience (Conner & Armitage, 2008).

### Research Problem

The aim of the presented study is to depict the risk of inaccuracy in the measurement of characteristics modelled in the structure of oppositional pairs and to answer the question of the role of such properties of a rating scale as: type, length, symmetry and anchoring, for shaping the measurement results of this kind of characteristics.

## Research Methodology

### Research General Background

The study was conducted as a quasi-experimental design, with random allocation of conditions and comparison of measurement results between groups and within the group.

### Participants

270 people participated in the study (including 204 women and 58 men; 8 people did not provide relevant information). The average age was 26 years (m=26.60, me=24, sd=7.34). The respondents were recruited from the student populations of two Polish universities, studying full-time or part-time, and majoring in pedagogy and psychology. The sample was randomly selected. The units of random selection were the students' class groups.

### Instruments and Procedures

In practice, it is difficult to select a suitably large sample of people displaying an insignificant and adequately high intensity of both components of a given characteristic. That is why, the research scenario predicted that prior to their replies, the respondents would imagine a given situation. This element of the scenario was based on the assumption that the conditions with which the respondents would identify will correspond to their choices of specific positions on rating scales.

For this reason, the respondents were instructed to determine the position on a rating scale, which would be chosen by a person characterised by a given attitude towards a freely selected object (person, thing, etc.). The research design was based on a two-group plan: 2 experimental conditions. In the first experimental condition, an ambivalent attitude was given, and in the second one – neutral. In each condition, 4 factors were taken into account that characterize the scale: scale type (discrete, continuous), length (5, 7, 9, 4, 6, 8, 100, 120, 140 points), symmetry of discrete scale (even, odd) and anchoring (bipolar, unipolar). However, the study was conducted as an incomplete design, leaving out unipolar discrete scales. It was aimed at preventing fatigue with the task, especially since the theory suggested more interesting and diverse results in the case of continuous scales than with discrete rating scales. The instruction for the ambivalence condition informed the respondents that their task consisted in determining the position on a rating scale which would be chosen by a person with a simultaneously positive and negative attitude, to the same degree. The instruction for the neutral condition informed that the task consisted in determining the position which would be selected by a person with an indifferent attitude, i.e., neither positive nor negative. Conditions were randomly assigned in such a manner that first the sheets were distributed by drawing lots, and then handed out according to a specific rule, which consisted in distributing them always starting with the person sitting closest to the entrance to the classroom, and then in a direction perpendicular to the board or screen. All the respondents were to use each scale in the questionnaire. As a result, data from 134 people responding to the ambivalence condition and 136 people responding to the neutral condition was collected.

The respondents were given access to the questionnaire which contained: bipolar even discrete scales, with 4, 6 and 8 points, bipolar odd discrete scales, with 5, 7 and 9 points, bipolar visual analogue scales, with the length of 100, 200 and 140 mm, and unipolar visual analogue scales, with the length of 100, 120 and 140 mm. The length of the scales corresponded to the most common length variants (Bandalos, 2018; Colman, Norris, & Preston, 1997). In the case of unipolar scales, the left pole was labelled with a value of 0% and the tight pole with 100%. In other instances, the scales were bipolar, which means that their left poles were labelled as maximum negative, and the right poles were labelled as maximum positive. In this case, the values closer to the right pole were treated as higher.

**Data Analysis**

During the analysis, the following tools were applied: descriptive statistics, Mann-Whitney U test to intergroup comparisons for independent groups, and

Friedmann non-parametric ANOVA for dependent groups. The latter was used in the assessment of intra-group differences between measurements made on different scales. Statistical hypotheses were verified with an assumed significance level of $\alpha=0.05$.

## Research Results

The decision to implement the Mann-Whitney U test was made due to the lack of assumed normal distribution of the compared variables and variance homogeneity in the compared groups. While it was possible to implement tests robust to the failure to fulfil the second assumption, the manner in which responses were given meant that only an extremely asymmetrical distribution of variables could be expected.

**Table 1.** Descriptive statistics for discrete scales

| Length of scale | Condition | Median | Quartile Deviation (QD) | Quartile 1 | Quartile 3 |
|---|---|---|---|---|---|
| 4 | N | 3.00 | 0.50 | 2.00 | 3.00 |
| | A | 3.00 | 0.50 | 2.00 | 3.00 |
| 6 | N | 4.00 | 0.50 | 3.00 | 4.00 |
| | A | 4.00 | 0.50 | 3.00 | 4.00 |
| 8 | N | 4.00 | 0.50 | 4.00 | 5.00 |
| | A | 5.00 | 0.50 | 4.00 | 5.00 |

N - neutral, A – ambivalent

In the case of bipolar odd scales, mid-values on rating scales (5-, 7-, 9-point) were selected regardless of conditions and without exception. On the other hand, visible differences occurred for even scales (Table 1). Here, the members of both groups chose slightly different values on the scales. What is also worth mentioning is inter-group diversity (QD), which did not occur for odd scales. A review of descriptive statistics indicates a generalised tendency to select values higher than the arithmetic mean of points on the scale. An exception was the 8-point scale, where the distribution of responses in the group with the neutral condition was right-skewed, whereas in the group with the ambivalent condition it was left-skewed. This would mean readiness to use lower instead of higher values on the even scale under the neutral condition, and higher instead of lower values under the ambivalent condition. Inter-

group comparison (Table 3) confirmed this result and also indicated something that the review of descriptive statistics does not reveal, i.e., that a difference similar to the one on the 8-point scale also occurred for the 4-point scale. Effect size for both the discussed scales is not large. However, it definitely indicates a connection between the condition and choice of value on the rating scale.

**Table 2.** Descriptive statistics for continuous scales

| Length of scale | Condition | Mean | Standard deviation | Median | Quartile 1 | Quartile 3 |
|---|---|---|---|---|---|---|
| 100 $_{b}$ | N | 49.68 | 2.51 | 50.00 | 48.00 | 51.00 |
| | A | 49.41 | 4.24 | 49.00 | 48.00 | 51.00 |
| 120 $_{b}$ | N | 60.38 | 2.75 | 60.00 | 59.00 | 63.00 |
| | A | 59.72 | 4.10 | 60.00 | 57.00 | 62.00 |
| 140 $_{b}$ | N | 69.25 | 3.24 | 69.00 | 67.00 | 71.00 |
| | A | 69.48 | 6.35 | 69.50 | 67.00 | 73.00 |
| 100 $_{u+}$ | N | 0.04 | 0.25 | 0.00 | 0.00 | 0.00 |
| | A | 48.66 | 16.37 | 50.00 | 47.27 | 53.00 |
| 100 $_{u-}$ | N | 0.04 | 0.29 | 0.00 | 0.00 | 0.00 |
| | A | 47.68 | 16.60 | 50.00 | 47.00 | 52.00 |
| 120 $_{u+}$ | N | 0.05 | 0.34 | 0.00 | 0.00 | 0.00 |
| | A | 58.94 | 18.24 | 61.00 | 58.00 | 64.00 |
| 120 $_{u-}$ | N | 0.06 | 0.40 | 0.00 | 0.00 | 0.00 |
| | A | 58.77 | 17.79 | 61.00 | 57.00 | 64.00 |
| 140 $_{u+}$ | N | 0.06 | 0.35 | 0.00 | 0.00 | 0.00 |
| | A | 67.97 | 19.32 | 71.00 | 66.00 | 74.00 |
| 140 $_{u-}$ | N | 0.04 | 0.29 | 0.00 | 0.00 | 0.00 |
| | A | 68.15 | 19.71 | 71.00 | 67.00 | 75.00 |

b – bipolar, u+ – unipolar positive, u- – unipolar negative

On the other hand, the analysis of data gathered through visual analogue scales indicates, similarly to odd discrete scales, a general tendency to choose the mid-point on a scale, regardless of the condition of providing responses and the length of the rating scale. A review of descriptive statistics (Table 2) indicates a slightly higher diversity in results in the group with the ambivalent condition. However, they are so insignificant (measurement with an accuracy of 1 mm), that even effect size (Table 3) for a scale with a length of 120 mm can hardly be taken for an

unquestionable justification for inter-group differences in the scope of choosing a position on the line.

**Table 3.** A comparison of results between the neutral (N) and ambivalent (A) conditions. The Mann-Whitney U test

| Length of scale | Sum of ranks (N) | Sum of ranks (A) | U | Z corr. | p | Effect size* |
|---|---|---|---|---|---|---|
| 5 | 18294.00 | 18291.00 | 8978.00 | -0.99 | 0.3200 | -0.06 |
| 7 | 18237.00 | 18348.00 | 8921.00 | -0.52 | 0.6014 | -0.03 |
| 9 | 18037.50 | 18547.50 | 8721.50 | -1.58 | 0.1136 | -0.10 |
| 4 | 17321.50 | 19263.50 | 8005.50 | -2.00 | 0.0459 | -0.12 |
| 6 | 17709.00 | 18876.00 | 8393.00 | -1.31 | 0.1895 | -0.08 |
| 8 | 17068.00 | 19517.00 | 7752.00 | -2.42 | 0.0157 | -0.15 |
| 100 | 18758.50 | 17826.50 | 8781.50 | 0.52 | 0.6041 | 0.03 |
| 120 | 19502.50 | 17082.50 | 8037.50 | 1.68 | 0.0924 | 0.10 |
| 140 | 18258.00 | 18327.00 | 8942.00 | -0.27 | 0.7908 | -0.02 |
| 100 $_{u+}$ | 9319.00 | 27266.00 | 3.00 | -15.14 | 0.0000 | -0.92 |
| 100 $_{u-}$ | 9318.00 | 27267.00 | 2.00 | -15.14 | 0.0000 | -0.92 |
| 120 $_{u+}$ | 9319.50 | 27265.50 | 3.50 | -15.14 | 0.0000 | -0.92 |
| 120 $_{u-}$ | 9320.00 | 27265.00 | 4.00 | -15.11 | 0.0000 | -0.92 |
| 140 $_{u+}$ | 9317.50 | 27267.50 | 1.50 | -15.12 | 0.0000 | -0.92 |
| 140 $_{u-}$ | 9316.00 | 27269.00 | 0.00 | -15.14 | 0.0000 | -0.92 |

* calculated according to the formula: $r=Z/\sqrt{n}$

A different situation occurs with respect to unipolar scales. According to expectations, the differences between conditions were definite and clear (Table 3). The persons with the neutral condition marked the left pole of both lines, precisely at the beginning. The persons with the ambivalent condition selected places which were similarly distant from the left pole of both lines, usually close to the middle. However, for the neutral condition, we noticed very high values of standard deviation in relation to arithmetic means, and extremely high positive coefficients of skewness (from 6.6 to 7.2). Both result from the occurrence of four observations, which clearly stand out from 0 on the rating scale – clearly in this case meaning approximately 2–3 mm.

To assess the role of scales length, previously all data was subjected to proportional transformation. The length of a scale proved to be important both in the

neutral (Friedman chi²=679.48, df=14, p=2.2e-16, effect size: Kendal's W=0.87) and ambivalent condition (Friedman chi²=1657.9, df=14, p=2.2e-16, effect size: Kendal's W=0.36), although in each according to a different scheme. In the first one, the measurements with continuous scales differed from the measurements with discrete scales, and also differences occurred due to the anchoring of the scale. In the second one, differences between discrete and continuous scales also occurred, while in almost all the comparisons of measurements, the bipolar continuous scales did not differ from the unipolar continuous scales.

## Discussion

The results of the conducted study are in accordance with the expectations that the properties of the rating scale play a role in determining the results of the measurement of characteristics modelled in the structure of oppositional pairs. They particularly coincide with the assumption and results obtained by other authors, that applying bipolar scales in the measurement of such characteristics decreases the accuracy of measurement. This consists in representing different states of a characteristic by the same value on a rating scale. However, bipolar scales are probably not completely insensitive to these differences, though in their case it may also be important whether they are discrete or sufficiently long. Such an assumption can be made following the result for discrete bipolar even scales, which was expressed by higher selection of lower values under the neutral condition, and higher values under the ambivalent condition. Even scales imply a search for a location corresponding with the intensity of a characteristic and, presumably, induce other behaviours in people with a neutral attitude than in people with an ambivalent one. The former choose lower values, and the latter higher. This interpretation, though seemingly justified, requires further study on even scales in the measurement of characteristics modelled in an oppositional structure.

In the case of bipolar visual scales, a similar result did not occur, though for the ambivalent condition the scales yielded more varied results (standard deviations) than for the neutral condition. They were clear and regular, though not significant enough to be recognised at the adopted confidence level. However, they allow for formulation of an assumption that low intensity of both tendencies has a more unequivocal representation on a bipolar rating scale than the intensity which corresponds to ambivalence, and that this representation constitutes the middle of the scale. This assumption is also supported by a significantly higher variation of results yielded by unipolar scales under the ambivalent condition than under

the neutral one, which suggests that ambivalence is less unequivocally identified with its corresponding position on the rating scale than neutrality.

The thesis concerning differences in the scope of the un-ambiguity of representation is immensely interesting, given the common and opposing standpoint on the significance of the mid-point of the bipolar rating scale. Naturally, the mid-point of the scale is still burdened by high risk of surjective modelling. In the presented study, under the ambivalent condition, the respondents were asked to mark places on unipolar scales which correspond with the same intensity of positive and negative attitude. Although a more distinct variation occurred in this scope than in the group under the neutral condition, the majority of the respondents determined positions close to the middle of the scale. Moreover, in contrast to the neutral condition, there was a large similarity between bipolar and unipolar visual measurements. On the other hand, this variation is still slight in comparison to the possibilities of choice offered by unipolar visual analogues scales. This in turn suggests that ambivalence can be generally associated with the mid-point of the rating scale, or those values of the scale which correspond with moderate intensity of opposite tendencies. Minimum and maximum, as well as values of quartile 1 and 3 indicate that under the ambivalence condition, low and high values of rating scales were selected less frequently. The thesis on the difference in the scope of un-ambiguity of representation creates space for a more subtle depiction of the issue of the mid-point of the scale and broadens the perspective on values close to the mid-point.

## Conclusions

The selection of a rating scale which is sufficient enough to lower the risk of surjective modelling, i.e., the possibility of identifying a given value with opposite psychological states, is significant for the accuracy and validity of measurement. Bipolar scales, insofar as there is a need to model the intensity of both tendencies, should be replaced by unipolar scales. In addition, of appropriate length. Even if bipolar even scales, with what is called forced choice, had the capacity to provide information which suggests differences in psychological states. In the presented study, such an effect was revealed in the form of increased selection of higher values on the scale under the ambivalent condition, and lower values under the neutral condition. A measurement of oppositional tendencies, which makes use of separate unipolar scales, provides more data.

**References**

Bandalos, D.L. (2018). *Measurement Theory and Applications for the Social Sciences*. New York, London: The Guliford Press.

Chadha, K.N. (2009). *Applied Psychometry*. Los Angeles–London–New Delhi–Singapore: Sage Publication.

Colman, A.M., Norris, C.E., & Preston, C.C. (1997). Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales. *Psychological Reports*, 80, 355–362.

Conner, M., & Armitage, C.J. (2008). Attitudinal Ambivalence. In W.D. Crano, & R. Prislin, *Attitudes and Attitude Change* (pp. 261–288). London, New York: Psychology Press Taylor & Francis Group.

DeVellis, R.F. (2017). *Scale Development. Theory and Application* (4 ed.). Thousand Oaks: Sage Publication.

Hadjiiski, L., Chan, H.-P., Sahiner, B., Helvie, M.A., & Roubidoux, M.A. (2007). Quasi-Continuous and Discrete Confidence Rating Scales for Observer Performance Studies: Effects on ROC Analysis. *Academic Radiology*, 14 (1), 38–48.

Preston, C.C., & Colman, A.M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104, 1–15.

Priester, J.R., & Petty, R.E. (1996). The Gradual Threshold Model of Ambivalence: Relating the Positive and Negative Bases of Attitudes to Subjective Ambivalence. *Journal of Personality and Social Psychology*, 71 (3), 431–449.

Svensson, E. (2000). Comparison of the Quality of Assessments Using Continuous and Discrete Ordinal Rating Scales. *Biometrical Journal*, 42 (4), 417–434.