The effect of shape: comparing different presentations of response

Botelho, Maria do Carmo (1st author)
Instituto Universitário de Lisboa (ISCTE-IUL), Unidade de Investigação em Desenvolvimento Empresarial (Unide-IUL)
Av Forças Armadas, s/n
1649-026 Lisboa, Portugal
E-mail: maria.botelho@iscte.pt

Calapez, Teresa (2nd author)
Instituto Universitário de Lisboa (ISCTE-IUL), Unidade de Investigação em Desenvolvimento Empresarial (Unide-IUL)
Av Forças Armadas, s/n
1649-026 Lisboa, Portugal
E-mail: teresa.calapez@iscte.pt

Ramos, Madalena (3rd author)
Instituto Universitário de Lisboa (ISCTE-IUL), Centro de Investigação e Estudos de Sociologia (CIES-IUL)
Av Forças Armadas, s/n
1649-026 Lisboa, Portugal
E-mail: madalena.ramos@iscte.pt

Attitudes and motivations, intrinsically subjective attributes, need often to be studied and evaluated in order to support decisions in different areas of knowledge. But how to evaluate these subjective, and in many aspects non-measurable, entities? Almost one hundred years ago, rating scales have been proposed, which intend to gather the "degree of affection" of an individual on a particular object or value. From then on, several studies have evaluated, compared and discussed the behaviour of different rating scales. The earlier discussions are related to the diversity of categories to be included in each variable - essentially on the number, but also on the inclusion of a neutral category (Green and Rao, 1970; Weng, 2004; Moors, 2008). Other studies focus on the adequacy of assumption of equal distances between adjacent categories implicit in the usual quantification (assigning consecutive integers to successive categories) (Green and Rao, 1970; Jamieson, 2004; Carifio, 2007).

Different presentations of response have been compared with classical univariate statistics. However, due to the fragility of the classical measure of location (mean) it seems appropriate a robust approach using several types of estimators less sensitive to extreme values and heavy distribution tails. The performance of robust estimators of location - applied to both quantitative and Likert-type variables - is presented in, among others Botelho (2008) and suggests that these type of estimators have a better performance in longer than in shorter items.

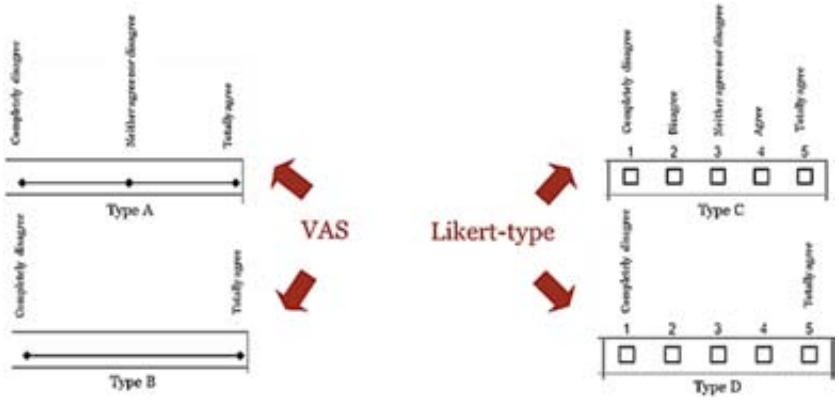
The aim of this field based work is to study in what measure different presentations of items induce different behaviours in scales distributions using both classical and robust approaches.

Participants and Procedures

The participants of this study were first year, first-time students, from several degree courses (Social Sciences, Management and Technological Sciences Courses) at a public university institution, whose course plan includes subjects from the scientific area of quantitative methods, defined as the target population. The sample consist in 727 participants, with age range between 16 and 56 years, the average age is 20.9 years (SD=6.7) and the most frequent age is 18 years old. The majority of

students are female (52.1%), with a Sciences' background from high school (71%) and had mathematics until their entrance in the university (91.4%). A questionnaire was applied to the referred population, in four versions. More specifically, the questionnaire included 18 items concerning three dimensions which are: a reworded subset of the Fennema-Sherman (1976) *Mathematics as a Male Domain* scale, the *Importance of understanding the concepts* scale proposed by Kloosterman and Stage (1992), and the *Usefulness of Mathematics* scale as modified by Kloosterman and Stage (1992). For each scale there are six statements. Half of the items included in each scale were written in a positive way and the other half in a negative one. Each item is a statement for which students should answer using a five-point concordance scale. In the four types of questionnaires the items were randomly assigned and in the same order. Responses were given differently: (i) Type A: using a 5cm long line, extreme-labelled, with a middle mark; (ii) Type B: using a 5cm long line, extreme-labelled, with no middle mark; (iii) Type C: using a five-point scale all labelled (completely disagree, disagree, do not agree nor disagree, and completely agree); and (iv) Type D: using five-point, extreme only labelled, Likert-type items (completely disagree and completely agree, respectively). The visual analogue scales and the Likert-type mentioned format for the items are represented in Figure 1. Some questions about individual

Figure 1: Items formats



characteristics such as gender, age or course, were also included. Items were randomly assigned and in the same order whatever the questionnaire type. Questionnaires were affected systematically in each class, so that an approximate number of each type was obtained, around 180 in each type. After the questionnaire application the group of participants were analyzed in their demographic characteristics and considered homogeneous in what concerns gender ($\chi^2_{(3)} = 5.385, p = 0.146$), age ($\chi^2_{(15)} = 14.815, p = 0.465$) and course field ($\chi^2_{(6)} = 0.303, p = 0.999$).

The aim of this investigation is to study in what measure different presentations of items (including "continuous" options, i.e. marking the option on a straight line, with or without a middle point and the use of all anchors *vs* extreme-only labels) induce different behaviours in scale distributions. The responses were compared at a scale level, using the constructed theoretically-defined summated scales.

Performance evaluation was undertaken in two steps. The first includes computation of several shape measures like skewness, kurtosis, normality and reliability indicators, as well as comparing them across response presentations, while in the second step location measures were used for point and interval estimation. Besides usual estimators such as the mean and the median, other types of estimators with a robust perspective were considered. Since the 1960's, the relevance of robustness in data analysis has been increasing, supported by a large theory development and growing computing capabilities, which allowed applications in real data. Tukey (Hoaglin et al, 1983) proposed a redescen-

dent M-estimator (Tukey's biweight) that uses different and smoothly decreasing weights as values shift from the distribution center.

In this work, besides Tukey's biweight, two other robust location estimators were considered, the first being the winsorized mean, where, in one or both tails according to data distribution, a predefined amount of observations, say $p \times 100\%$, is replaced by the sample's p^{th} quantile ($(\frac{p}{2})^{th}$ and $(1-\frac{p}{2})^{th}$, respectively, if in both tails) (Keselman et al, 2002). The last robust estimator considered, the LTS estimator, is simpler and easier to compute than M-estimators and was chosen due to the fact that uses rank values, thus an appropriate choice to accommodate ordinal data or low variability of the data. To calculate the LTS estimate consider a random sample of size n, and $n-h+1$ subsamples with h ordered observations. The LTS estimate corresponds to the mean of the the subsample with smallest associated sum of squares, i.e. the subsample mean with a greatest values concentration around that mean (Rousseeuw e Leroy, 1987). The LTS estimator can reveal the response concentration in each scale thus constituting an interesting measure to compare the influence of response types.

In order to compare each scale's location measures across questionnaires types, both point estimates and bootstrap confidence interval estimates were computed for the mean, the median and each of the three robust estimators considered. Bootstrap confidence intervals estimates can be constructed using several methods, the more obvious being the percentile method. This procedure, proposed by Lunneborg (2000), is based on a known T_n or on the correspondent estimate. When it is not possible to find the exact bootstrap distribution, T_n, a Monte Carlo's method is used in order to obtain an approximate distribution. First, the B bootstrap replications, generated by the T_n statistic applied to each of the B bootstrap samples, are sorted increasingly: $T^*_{1:B} \leq T^*_{2:B} \leq \ldots \leq T^*_{B:B}$. The approximate distribution is

$$R^B_{n,Boot}(T^*_{j:B}) = B^{-1} \sum_{b=1}^{B} I(T^*_{n,b} \leq T^*_{j:B}) = j/B, \quad (1)$$

for $j = 1, 2, \ldots, B$. Thus, the order percentiles $\alpha/2$ and $(1-\alpha/2)$ are estimated, respectively, by $Q^*(\alpha/2) \cong T^*_{j_1:B}$ and $Q^*(1-\alpha/2) \cong T^*_{j_2:B}$, where $j_1 = (B\alpha/2)+1$ and $j_2 = [B(1-\alpha/2)]+1$. For $\alpha = 0.05$, $T^*_{j_1:B}$ and $T^*_{j_2:B}$ estimate the 0.025 and 0.975 percentiles so $(T^*_{j_1:B}, T^*_{j_2:B})$ is a bilateral 95% confidence interval for θ parameter.

Wilcox (2003) refers that, when applied with M-estimators, the percentile method for constructing confidence intervals gives better results than others.

Results

The scale's distribution shape measures across questionnaires types were compared. The distributions are clearly asymmetric, with decreasing values of skewness for the three scales in study. The *Mathematics as a Male Domain* scale has a more asymmetric distribution than the *Understanding the Concepts is Important in Mathematics* scale and reveals also a greater number of outliers. The third scale, *Usefulness of Mathematics*, has distributions with values of skewness and kurtosis slightly lower than the previous two and fewer outliers. Within each scale, results are not substantially different across questionnaire types, especially in what skewness is concerned. All but one distributions show lack of normality (Table 1), the exception being the *Usefulness of Mathematics* scale computed over type D questionnaires (5-point Likert-type items, labelled only at the extremes).

Computed Cronbach's α are generally high, and in line with results from previous studies (Fennema and Sherman, 1976; Kloosterman and Stage, 1992). In all scales, whatever the item format, similar and high consistency was found, except for the continuous formats of the *Understanding the concepts* scale, which revealed low consistency.

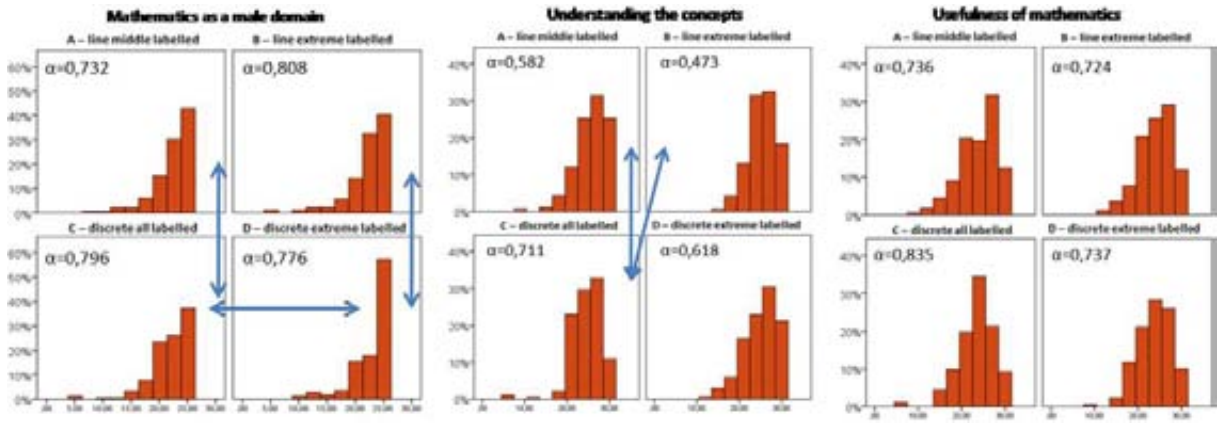
In more asymmetric distributions, different types of questionnaires seem to lead to more differences in response distribution (Figure 2). In the less asymmetric distributions - from the *Usefulness of*

Table 1: Distribution shape indicators.

Scales	Quest. types	Shape				Normality			
		Skweness		Kurtosis		D'Agostino & Pearson		Jarque & Bera	
		g1	p-value	g2	p-value	test	p-value	test	p-value
Mathematics as a male domain (higher value: No; 5 items)	A	-1.928	0.000	4.254	0.000	79.797	0.000	233.261	0.000
	B	-2.322	0.000	6.582	0.000	100.580	0.000	446.877	0.000
	C	-2.028	0.000	6.835	0.000	95.689	0.000	457.234	0.000
	D	-1.988	0.000	3.917	0.000	78.732	0.000	215.777	0.000
Understanding the concepts	A	-1.157	0.000	2.507	0.000	40.414	0.000	76.064	0.000
	B	-0.552	0.004	-0.164	0.737	8.225	0.016	8.682	0.013
	C	-1.722	0.000	7.025	0.000	85.365	0.000	441.769	0.000
	D	-0.772	0.000	0.028	0.814	14.755	0.001	16.704	0.000
Usefulness of mathematics	A	-0.725	0.000	0.218	0.469	14.179	0.001	15.485	0.000
	B	-0.535	0.006	-0.216	0.611	7.896	0.019	8.292	0.016
	C	-0.969	0.000	2.577	0.000	37.456	0.000	78.843	0.000
	D	-0.405	0.032	-0.024	0.928	4.632	0.099	4.590	0.101

Mathematics scale - the Kruskal-Wallis test did not reveal significant differences across questionnaire types.

Figure 2: Response distributions and reliability of scales.



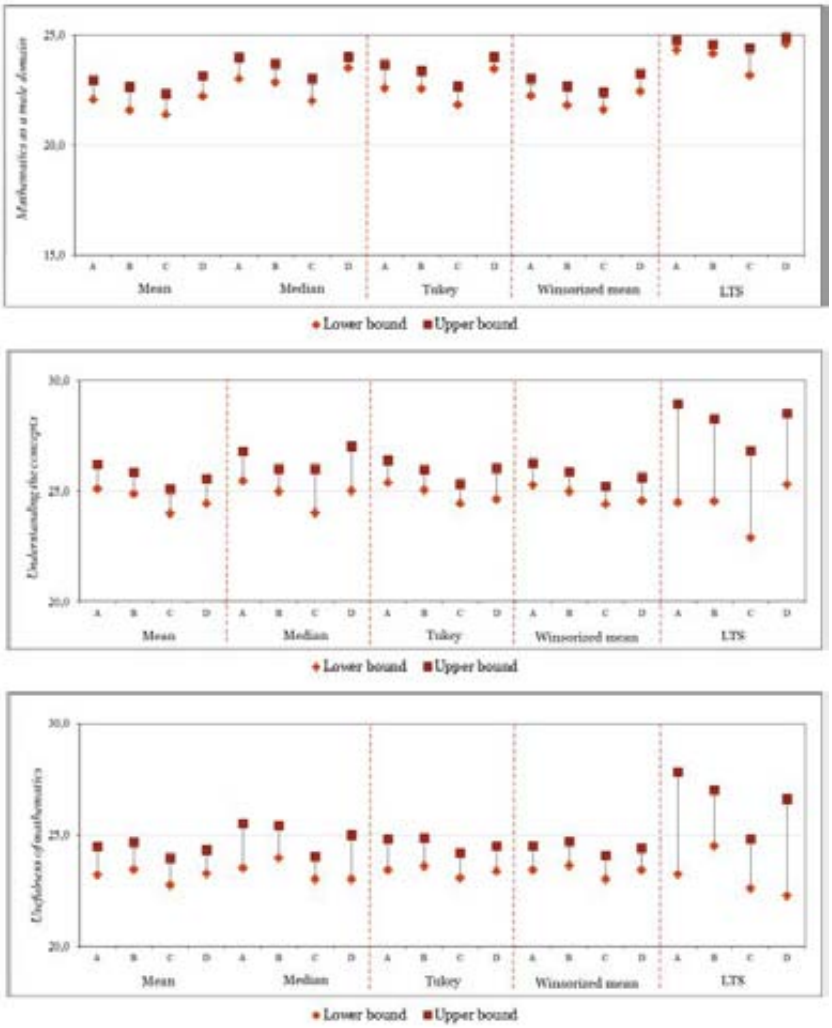
All five location point estimates, for each of the three scales and four questionnaire types are presented in Table 2. Type C questionnaires (5-point, all labelled Likert-type items) reveal lower location estimates for all but one of the elected estimators, the exception being the LTS estimate in the *Understanding the Concepts is Important in Mathematics* scale. In the more asymmetric distributions, obtained for the *Mathematics as a Male Domain* scale, type D questionnaires (5-point Likert-type items, labelled only at the extremes), reveal a higher concentration of responses in the right tail of the distribution, based on LTS values. As for the *Understanding the Concepts is Important in Mathematics* scale, which display lower asymmetry, distributions from questionnaire types A and D (the ones with labels only at the extremes, the former being continuous and the latter discrete) have a higher concentration of responses in the right tail. In the less asymmetric scale, *Usefulness of Mathematics*, the higher concentration in the right tail occurs in the continuous formats, types A and B.

In Figure 3 the bootstrap confidence intervals are presented. Results show a similar profile for the several estimators among the different types of questionnaires, except for the LTS, where the confidence intervals are wider when applied to distributions with less asymmetry. Median and Tukey's biweight estimators have a clear similarity when applied to the most asymmetric distributions.

Table 2: Location estimates.

Scales	Quest. types	Location estimators				
		Mean	Median	Tukey's biweight	Winsorized mean	LTS
Mathematics as a male domain (higher value: No; 5 items)	A	22.49	23.44	23.11	22.65	24.59
	B	22.13	23.32	23.00	22.30	24.39
	C	21.84	23.00	22.23	22.02	24.16
	D	22.71	24.00	23.74	22.83	24.74
Understanding the concepts	A	25.63	26.00	25.93	25.75	28.46
	B	25.35	25.68	25.47	25.42	25.19
	C	24.56	25.00	24.86	24.80	25.74
	D	25.02	26.00	25.47	25.09	28.08
Usefulness of mathematics	A	23.85	24.24	24.10	23.99	26.33
	B	24.07	24.64	24.22	24.16	26.07
	C	23.36	24.00	23.57	23.49	23.92
	D	23.80	24.00	23.88	23.91	24.34

Figure 3: Bootstrap confidence intervals.



For each estimator used, all formats of response produce similar and overlapping confidence intervals, except for type C (5-point, all labelled Likert-type items) in the most asymmetric scales. Confidence intervals derived from the robust estimates Tukey's biweight and LTS reveal lower values for both limits.

In the distributions with less asymmetry type C format of responses leads to confidence intervals always with slightly lower limits, whatever the estimator used. Differences are more clear between types C and D, i.e. between the discrete formats. In continuous formats the confidence intervals show an identical profile whatever the level of asymmetry.

Conclusions

The behaviour of the scales corresponding to the four types of response formats was not exactly the same. In the *Mathematics as a Male Domain* scale, a severe asymmetry and extreme values was found; some asymmetry and fewer extreme values in the *Understanding Concepts* scale; a slight asymmetry in the *Usefulness of Mathematics* scale. In general, all scales have shown high consistency, except for the *Understanding Concepts* scale with continuous formats. Different response formats lead to differences in the distributions of responses, particularly in the more asymmetric scales. The discrete format with all points labelled show a different behaviour in all scales when compared to continuous format or discrete with only the extremes labelled. In the discrete formats the responses seems to "follow the labels": when only the extreme are labelled, the estimate is more concentrated in higher values. The discrete format with only the extremes labelled shows similarity to continuous formats in the intervals estimates. Thus, the use of statistic measures designed for metric variables in Likert-type items seems to be more adequate if labels are used only at the extremes.

REFERENCES (RÉFERENCES)

Botelho, M.C. (2008), Técnicas Robustas de Estimação. Amostragem, variáveis e dimensões, PhD dissertation, Lisbon University Institute (ISCTE-IUL), Lisbon.

Carifio, J. and Perla, J.R. (2007) Ten Common Misunderstandings, Misconceptions, Persistent Myths and Urban Legends about Likert Scales and Likert Response Formats and their Antidotes, *Journal of Social Sciences* 3 (3), 106-116.

Fennema, E.H. and Sherman, J.A. (1976). Fennema-Sherman mathematics attitudes scales: instrument designed to measure attitudes toward mathematics. *Journal for Research in Mathematics Education*, 7(5), 324-326.

Green, P.E. and Rao, V.R. (1970) Rating Scales and Information Recovery - How many scales and response categories to use?, *Journal of Marketing*, 34 (Jul 1970), 33-39.

Hoaglin, D.C., Mosteller, F. and Tukey, J.W. (1983), *Understanding robust and exploratory data analysis*, John Wiley & Sons.

Jamieson, S. (2004) Likert scales: how to (ab)use them, *Medical Education*, 38, 1212-1218.

Keselman, H.J., Wilcox, R.R, Othman, A.R. and Fradette, K. (2002), Trimming, Transforming Statistics, And Bootstrapping: Circumventing the Biasing Effects Of Heterescedasticity And Nonnormality, *Journal of Modern Applied Statistical Methods*, 1(2): 288-309.

Kloosterman, P. and Stage, F.K. (1992). Measuring beliefs about mathematical problem solving. *School Science and Mathematics*, 92, 109-115.

Lunneborg, C.E. (2000), *Data analysis by resampling: concepts and applications*, (Pacic Grove, CA, Duxbury).

Moors, G. (2008) Exploring the effect of a middle response category on response style in attitude measurement, *Quality & Quantity*, 42, 779-794.

Rousseeuw, P.J. and Leroy, A.M. (1987) *Robust Regression and Outlier Detection*, John Wiley & Sons.

Weng, L. (2004) Impact Of The Number Of Response Categories And Anchor Labels On Coefficient Alpha And Test-Retest Reliability, *Educational and Psychological Measurement*, 64(6), 956-972.

Wilcox, R.R. (2003), *Applying Contemporary Statistical Techniques*, Academic Press.