Running Head: Children's responses to rating scales

The effects of changes in the order of verbal labels and numerical values on children's scores on attitude and rating scales

Lucy Betts & James Harltey

Abstract

Research with adults has shown that variations in verbal labels and numerical scale values on rating scales can affect the responses given. However, few studies have been conducted with children. The study aimed to examine potential differences in children's responses to Likert-type rating scales according to their anchor points and scale direction, and to see whether or not such differences were stable over time. 130 British children, aged 9 to 11, completed six sets of Likert-type rating scales, presented in four different ways varying the position of positive labels and numerical values. The results showed, both initially and 8-12 weeks later, that presenting a positive label or a high score on the left of a scale led to significantly higher mean scores than did the other variations. These findings indicate that different arrangements of rating scales can produce different results which has clear implications for the administration of scales with children.

Key words: Likert scales; rating scales; questionnaires; response formats; research design; children.

The effects of changes in the order of verbal labels and numerical values on children's scores on attitude and rating scales

Considerable research has been conducted on the effects of the format and the wording of rating scales with adults but relatively little has been done in this respect with children. This is surprising, given that self-report measures and questionnaires are widely used by researchers and clinicians with children. It is also important that the data generated is both valid and reliable when questionnaires are developed and used with children (see Bell, 2007).

Work with adults has focused on the effects of the number of items in a scale, the presence or absence of neutral points, the wording and direction of the scale items (from positive to negative and vice versa), as well as whether or not the scale ratings should go from low to high, high to low or, in some cases, to and from a neutral point. Examples of this literature can be found in the many papers of Dillman, (e.g., Dillman & Christian, 2005), Schwartz (e.g., Schwartz & Oyserman, 2001), and Tourangeau (e.g., Tourangeau, Couper, & Conrad, 2004) and more recent overview studies have been published (e.g., see Hartley & Betts, 2010, Shulruf, Hattie & Dixon, 2008, and Van Schaik and Ling, 2007).

It is important, therefore, to see if these effects are found with children. Although there are far fewer studies with children, the format of rating scales has been discussed in studies when children have been participants (e.g., see Beitchman & Corradini, 1988; Fuchs, 2005). Thus there have been studies with children on the effects of using different numbers of items in a scale (e.g., Borgers, Hox, & Sikkel, 2004; Chambers & Johnston, 2002; Pajares, Hartley, & Valiente, 2001), on the effects of including a neutral point in a scale (e.g., Raaijamakers, Van Hoof, Hart, Verbogt, & Volleberg, 2000), and on the particular wording of the scales (e.g., Borgers et al.,

2004; Marsh, 1986; Pantell & Lewis, 1987). The findings of these studies suggest that the response formats of children's questionnaires not only have the potential to influence the psychometric properties of the questionnaire but also the quality of the data (Bell, 2007). In addition Bell (2007) argues that response formats for children should be selected carefully to avoid confusion, boredom, and cognitive overload.

Nonetheless, it is not clear from these studies whether or not the performance of children is different from that of adults. Chambers and Johnson (2002) suggest that children between 5 and 6 years old make more extreme judgements than do older children and adults, and some investigators have adjusted their scales to make them easier for children to use. Lau and Lee (2008), for example, changed a scale with 7 points to one with 5 points on the basis that the 10-17 year old respondents in their sample were younger than the ones used in their original study with adults. Borgers et al. (2004) also recommend offering 'about four response options' (p. 30) when the participants are aged 8-16, and Borgers and Hox (2001) recommend offering four to five response options, randomising the order of item presentation, and not using ambiguous response scales to avoid non-response when working with children. Other investigators have used oral presentations, and very young children have often been asked to respond by pointing to particular options, sometimes presented in the form of 'smiley-faces' or some other type of visual display (e.g., see Chambers & Johnston, 2002; Davis, Leman, & Barrett, 2007; Lockl & Schneider, 2002; Teucher, 2008).

However, many investigators employing children as respondents do not appear to have considered these features very important. For example, Corcoran and Fischer (2000) have provided a compendium of 49 published scales for children that cover a wide range of topics (e.g., assertiveness, loneliness, self-control, self-esteem, etc.). Inspection of these scales shows that (i) 27 of them are Likert-type scales, (ii)

21 of these have the negative end of the scale on the left, and (iii) although there are 3, 4, 5, 6, 7 and 8 point scales, 11 of them use the 5 point scale response format. Also, in this connection, we note with some surprise that *none* of the scales included by Corcoran and Fischer (2000) have the highest score value on the left, unlike some scales with adults. All of them, therefore, start with a scale value of 0 or 1.

Such considerations have implications that move beyond the research arena and into the applied or clinical arena when children are asked to report their experiences. For example, when they self-report pain (Stanford, Chambers, & Craig, 2006; Stinson, Kavanagh, Yamada, Gill, & Bonnie, 2006), somatic complaints (Jellesma, Rieffe, & Terwogt, 2007), or psychopathic traits (van Baardewijk et al., 2006).

In this paper we focus on the effects of making one fairly common design decision with respect to the format of typical rating scales for work with children that, so far, seems to have escaped attention. In brief we were interested to see what would happen when we varied both the position of the labels and the order of the sequence of the scale points on a set of Likert-type scales. Accordingly, we decided that we would compare the following four schematic possibilities:

Very much 6 – 5 – 4 – 3 – 2 – 1 Not at all

(positive wording and high rating on the left)

Very much 1 – 2 – 3 – 4 – 5 – 6 Not at all

(positive wording and low rating on the left)

Not at all 6 – 5 – 4 – 3 – 2 – 1 Very much

(negative wording and high rating on the left)

Not at all 1 – 2 – 3 – 4 – 5 – 6 Very much

(negative wording and low rating on the left)

Our question of interest was: Would any one of these questionnaire formats produce higher mean scale scores than any of the others?

Previous research with adults would suggest that we might expect higher mean scores on the first of these scales, that is the one with positive wording and high ratings on the left (e.g., see Hartley & Betts, 2010; Nicholls, Orr, Okubo, & Loftus, 2006; and Weems, Onwuegbuzie, Schreiber, & Eggers, 2003) but, to date, this particular question does not appear to have been asked with respect to children**.** Furthermore, in the light of the literature described above, we wondered whether or not children might make less differentiated responses than appears to be the case with adults because of their more limited cognitive abilities (Beitchman & Corradini, 1988; Fuchs, 2005).

Method

The study was part of a larger study designed to investigate children's peer relationships, school adjustment, and psychosocial adjustment. Pupils from five British schools in a range of socio-economic areas took part. The head teachers of the schools concerned gave their consent for the study to take place, and the parents of all of the children who were in the target classrooms were informed of the study by letter and given the option of withdrawing their children from it. The children were also asked to give their verbal consent. The study took place in two stages.

At Time 1 we recruited one hundred and eighty seven school children (aged 9 – 11, 89 males; 98 females; 10 gender unspecified) from 12 classrooms. The sample comprised those children whose parents did not withdraw their children from the sample, 19 eligible children did not participate because their parents removed them from the sample. The first author distributed in each class a set of five scenarios concerned with peer influence (Burton, Ray, & Mehta, 2003). The scenarios were structured in the same way such that the same two children completed a class test where one of the children witnessed (target child) the other child cheating. The only difference across the scenarios was the nature of the peer influence which was manipulated by changing the ending of the scenarios (see Appendix 1).

The researcher read out the first scenario to the children who, after listening to it, were required to complete six rating scales concerning the target child's actions (see Appendix 2). Three of the rating scale items were positively worded (items 2, 4, and 6) and three of the rating scale items were negatively worded (items 1, 3, and 5). Four versions of the questionnaire were developed to reflect the scale formats described above. When all of the children had completed their ratings, the researcher then read out the text for scenario two, and this procedure was repeated with the children using the same scale format throughout the study. (The text in Appendix 2 provides an example of how this worked in practice for the *first* of our four scale formats, i.e. scales with the positive wordings and highest ratings on the left). The whole procedure was repeated again two to three months later (at Time 2), with each child completing the questionnaires in the same format and order as they had at Time 1. The questionnaires were matched through creating a unique identifier for each participant.

Data were recorded from each child who was present at Time 1 and 2, but only data available from children present on both occasions were used in the statistical analyses. The sample size was reduced to 130 (60 male, 64 female), as some participants did not complete all of the items at both times (n = 29), whereas others did not complete the scale at Time 2 (n = 38).

In all cases, negatively worded scale items for the scenarios were reverse-scored for the purpose of analysis (e.g. see Appendix 2, items 1, 3 and 5).

Results

Table 1 shows the descriptive statistics obtained at Time 1 and Time 2 for each scenario. These data were analysed using a four-way ANOVA: 2 (anchor word label: left 'very…' versus left 'not at all…') x 2 (scale value: left = 1 versus left = 6) x 2 (Time: Time 1 versus Time 2) x 5 (scenario). Time and scenario were repeated measures. This analysis allowed us to control for the possible effects of age, time, and scenario on the results. The ANOVA summary Table is shown Table 2.

[Insert Table1 and Table 2 about here]

There was a significant interaction between time, anchor word, and scale. To explore the interaction, simple effects were calculated. There was a simple effect of anchor word at Time 1, $F(1, 207.54) = 62.10$, $p < .001$, and at Time 2, $F(1, 207.54) = 40.33$, $p < .001$. At both Times, regardless of scenario, children awarded higher scores when left = very ($M_{\text{Time 1}} = 3.59$, $SE_{\text{Time 1}} = .06$, $M_{\text{Time 2}} = 3.55$, $SE_{\text{Time 2}} = .07$) than when left = not ($M_{\text{Time1}} = 3.17$, $SD_{\text{Time 1}} = .06$, $M_{\text{Time 2}} = 3.20$, $SE_{\text{Time 2}} = .06$). There was also a simple effect of scale at Time 1, $F(1, 207.54) = 12.12$, $p < .001$, and at Time 2, $F(1, 207.54) = 4.66$, $p < .05$. At Time 1, regardless of scenario, children

awarded higher scores when left = 6 ($M_{\text{Time 1}}$= 3.47, $SE_{\text{Time 1}}$= .06, $M_{\text{Time 2}}$= 3.55, $SE_{\text{Time 2}}$= .07) than when left = 1 (M_{Time1} = 3.29, $SE_{\text{Time 1}}$= .06, $M_{\text{Time 2}}$ = 3.32, $SE_{\text{Time 2}}$ = .06).

The analysis was repeated to explore potential age effects across the scale response format. When age was controlled for, there was no significant difference between the scores obtained for 9-, 10- and 11-year-olds and there were no significant interactions between the variables tested and the age of the children.

We next compared the scores on each scenario of those children who completed both testing sessions with those who did only one, to check whether or not the children who completed both test sessions were different in any way from those who completed only one. This analysis showed that there were no significant differences between the scores obtained for those children who completed both data sets, and those who did not and there were no significant interactions between the variables tested and whether or not the children had taken part.

Finally, we calculated Cronbach's alpha values for each of the scenarios at Time 1 and Time 2. Table 3 shows the results. It can be seen that the internal consistency of the scales ranged from low to modest (the implications of this will be considered in the discussion).

[Insert Table 3 about here]

So, to recapitulate, the main finding was, when all else was controlled for, that higher ratings were obtained when the anchor word on the left was positive (e.g.'very …') for all of the scenarios. Similarly, higher scores were obtained when the scale direction was left = 6. These findings, however, must be treated with caution

because, as shown in Table 3, the analyses of the internal consistencies of the scales and the confidence intervals for each scenario revealed that these were low to modest, at best. Perhaps, in these circumstances, it is more appropriate to use a more stringent significance value of $p < .01$. If we adopt this procedure here then our main result still stand, with exception of the simple effect of scale direction at Time 1.

Discussion

We argued in the Introduction that different investigators have used a variety of formats for Likert-type scales in their research with children, but that few of them have examined the effects of these different formats on the results obtained. Most, we think, as noted in the Introduction, have used scales with a 'standard format' with both negative wording and a scale point of 0 or 1 on the left.

However, the results obtained in our enquiry showed that children using scales with *positive* wording on the left consistently produced higher total scale scores than did children using the more typical settings, as did the scales with the *higher* value anchor points on the left. Our findings thus replicate some of the results reported earlier with adult participants (e.g, see Nicholls et al., 2006; and Weems et al., 2003) and in particular Hartley and Betts (2010). However, the results of this current study with children were less differentiated than were the previous ones with adults. In the adult study the highest ratings were given to scales that had both the positive label *and* the highest scale value on the left. In the present paper - with children - the highest ratings went to scales with the positive wording on the left irrespective of the scale value.

Why there should be a bias towards the left – both for adults and for children is a matter of conjecture. Three explanations that have been proffered for this phenomenon are: (i) the 'pseudo-neglect' effect, i.e. an attentional bias in normal

individuals that makes the left-sided features of a stimulus more salient than those on the right (Nicholls et al., 2006), (ii) a habitual left-to-right reading effect in native speakers of English (e.g., see Dolnicar & Grun, 2007), and (iii) a cognitive expectation effect, where again people expect to find larger numbers and positive words more on the left than the right (e.g., see Holbrook, Krosnick, Carson & Mitchell, 2000; Rozin, Berman & Royzman, 2010; Tourangeau et al., 2004). All three of these possible explanations invoke a visual attentional bias towards the left. Studies with eye-tracking devices involving readers with different nationalities might help resolve this issue (e.g., see Radach, Huestegge, & Reilly, 2008; and Shrestha & Lenz, 2007).

An additional explanation of the findings, and of how children use questionnaires, may reside in their more limited cognitive abilities. Specifically, as noted in the Introduction, differences in children's ages and cognitive abilities may influence their understanding of scale items, anchor points, and response scales (Beitchman & Corradini, 1988; Fuchs, 2005). However, when we tested this idea by repeating the analyses described above, but this time controlling for age, we found no differential effects for respondents aged 9, 10 and 11 years. Similarly, we found no significant differences between the data obtained at the two time periods, Time 1 and Time 2, although of course this was over the relatively short period of two to three months. Nonetheless, it would be wise if, in future, researchers and practitioners consider further the importance of children's cognitive abilities for scale completion and for the stability of responses in longitudinal studies and repeated clinical assessments.

Our main conclusion - that children respond differently to different rating scale formats - is, of course, going beyond the data reported here in the sense that the

findings are limited to this particular enquiry. We recognise that our results represent children's responses to particular six-point scales and, that there were some negatively- and some positively-worded items within them. Another aspect of these findings is that they result from making forced responses as there was no neutral response, and that this may be evidenced in the low internal consistency of the scales (as shown in Table 3). Indeed, such similar low consistencies seem to be a feature of work with children. For example, the *Pictorial Scale of Perceived Competence and Social Acceptance* has internal consistencies between .48 and .88 (Mantzicopoulos, French, & Maller, 2004); and Davis-Kean and Sandler (2001) reported a similar range of results in a meta-analysis of children's self-esteem measures. These low to modest internal consistencies may have emerged in the present study because of the relatively small number of items in the scale and because the scales may have been multifactorial, with different items assessing different facets of peer influence. Consequently, the low to modest internal consistencies may have reduced the power of the study.

Alternatively, such low to moderate internal consistencies may reflect the children's developing cognitive abilities (Arseneault, Kim-Cohen, Taylor, Caspi, & Moffitt, 2005). Perhaps phenomena such as these can be better studied through such children's self-reports (Stormshak & Webster-Stratton, 1999; Sturgess, Rodger, & Ozanne, 2002). Further research is needed to address these issues, and more systematic work needs to be done to study the effects of different cognitive tasks demanded by different wording and scale values, different age groups and different topics.

The main import of our paper, however, is to suggest that researchers consider factors such as these when planning their studies and when reviewing the results of

earlier findings, especially when different formats are used in the same investigation. Indeed it is crucial for researchers to take into consideration how children engage with the many different scale formats available so that they collect the most accurate data possible. For example, when using several questionnaires, researchers may want to consider using the same response format for each one, as variations in response format may confound the results. Specifically, reducing variations of response formats within the same study would reduce potential bias and prevent artificially increases in mean scores because of the questionnaire format. Similarly, practitioners need to consider the format of the questionnaires and rating scales that they provide for children, as it may be that asking children to complete multiple scales with different response formats could confound the accuracy of the data.

References

Arseneault, L., Kim-Cohen, J., Taylor, A., Caspi, A. & Moffitt, T. E. (2005) Psychometric evaluation of 5- and 7- year-old children's self-reports of conduct problems, *Journal of Abnormal Child Psychology,* 33, 537-550.

Barnette, J. J. (2000) Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded items, *Educational & Psychological Measurement,* 60(3), 361-370.

Beitchman, J. H. & Corradini, A. (1988) Self-report measures for use with children: A review and comment, *Journal of Clinical Psychology*, 44, 477-490.

Bell, A. (2007). Designing and testing questionnaires for children. *Journal of Research in Nursing, 12,* 461-469.

Borgers, N. & Hox, J. (2001) Item nonresponse in questionnaire research with children, *Journal of Official Statistics,* 71, 321-335.

Borgers, N., Hox, J. & Sikkel, D. (2004) Response effects in surveys on children and adolescents: The effects of number of response options, negative wording, and neutral-midpoint, *Quality and Quantity,* 38(1), 17-33.

Burton, B. A., Ray, G. E. & Mehta, S. (2003) Children's evaluations of peer influence: The role of relationship type and social situation, *Child Study Journal,* 33(4), 235-255.

Chambers, C. T. & Johnston, C. (2002) Developmental differences in children's use of rating scales, *Journal of Pediatric Psychology, 27,* 27-36.

Corcoran, K. & Fischer, J. (2000) *Measures for clinical practice, vol. 1* (3rd ed New York, Free Press).

Davis-Kean, P. E. & Sandler, H. M. (2001) A meta-analysis of self-esteem for young children: A framework for future measures, *Child Development,* 72, 88-906.

Davis, S. C., Leman, P. J. & Barrett, M. (2007) Children's implicit and explicit ethnic group attitudes, ethnic group identification, and self-esteem, *International Journal of Behavioral Development*, 31, 514-525.

Dillman, D. & Christian, L. (2005) Survey mode as a source of instability in responses across surveys, *Field Methods,* 17, 30-52.

Dolnicar, S. & Grun, B. (2007) Cross cultural differences in survey response patterns. *International Marketing Review,* 24(2), 127-143.

Fuchs, M. (2005) Children and adolescents as respondents. Experiments on question order, response order, scale effects and the effect of numeric values associated with response options, *Journal of Official Statistics*, 21, 701-725.

Hartley, J. & Betts, L. (2010) Four layouts and a finding: The effects of changes in the order of the verbal labels and the numerical values on Likert-type scale scores, *International Journal of Social Research Methodology,* 13, 17-27.

Holbrook, A., Krosnick, J. A., Carson, T. T. & Mitchell, R. C. (2000) Violating conversational conventions disrupts cognitive processing of attitude questionnaires, *Journal of Experimental Social Psychology,* 36, 465-494.

Jellesma, F. C., Rieffe, C. & Terwogt, M. M. (2007) The Somantic Complaint List: Validation of a self-report questionnaire assessing somantic complaints in children, *Journal of Psychosomatic Research,* 63, 399-401.

Lau, K-L. & Lee, J. C. K. (2008) Validation of a Chinese goal orientation questionnaire, *British Journal of Educational Psychology,* 78, 331-353.

Lockl, K. & Schnieder, W. (2002) Developmental trends in children's feeling-of-knowing judgements, *International Journal of Behavioral Development,* 26(4), 327-333.

Mantzicopoulos, P., French, B. F & Maller, S. J. (2004) Factor structure of the pictorial scale of perceived competence and social acceptance with two pre-elementary samples, *Child Development,* 75, 1214-1228.

Marsh, H. W. (1986) Negative item bias in rating scales for preadolescent children: A cognitive-developmental phenomenon, *Developmental Psychology,* 22, 37-49.

Nicholls, M. R., Orr, C. A., Okubo, M. & Loftus, A. (2006) Satisfaction guaranteed: The effect of spatial biases on responses to Likert scales, *Psychological Science,* 17, 1027-1028.

Pantell, R. H. & Lewis, C. C. (1987) Measuring the impact of medical care on children, *Journal of Chronic Diseases,* 40, Supplement 1: 99S – 108S.

Pajares, F., Hartley, J. & Valiante, G. (2001) Response format in writing self-efficacy assessment: Greater discrimination increases prediction, *Measurement & Evaluation in Counseling & Development,* 33, 214-221.

Raaijmakers, Q.A.W., Van Hoof, A., Hart, H. T., Verbogt, T. F. M. A. & Volleberg, W. A. M. (2000) Adolescents' midpoint responses on Likert-type scale items: Neutral or missing values? *International Journal of Public Opinion Research,* 12, 209-217.

Radach, R., Huestegge, L. & Reilly, R. (2008) The role of global top-down factors in local eye-movement control in reading, *Psychological Research,* 72, 675-688.

Rozin, P., Berman, L. & Royzman, E. (2010) Biases in use of positive and negative words across twenty natural languages, *Cognition and Emotion,* 24(3), 536-548.

Schwarz, N. & Oyserman, D. (2001) Asking questions about behavior: Cognition, communication, and questionnaire construction, *American Journal of Evaluation*, 22, 127-160.

Shrestha, S. & Lenz, K. (2007) Eye gaze patterns whilst searching vs browsing a website, *Usability News*, 9(1), Retrieved October 25, 2008, from http://psychology.wichita.edu/surl/usabilitynews/91/eyegaze.asp

Shulruf, B., Hattie, J. & Dixon, R. (2008) Factors affecting responses to Likert type questionnaires: Introduction of the ImpExp, a new comprehensive model, *Social Psychology of Education*, 11, 59-78.

Stanford, E. A., Chambers, C. T. & Craig, K. D. (2006) The role of developmental factors in predicting young children's use of self-report scales for pain, *Pain*, 120, 16-23.

Stinson, J. N., Kavanagh, T., Yamada, J., Gill, N. & Bonnie, S. (2006) Systematic review of the psychometric properties, interpretability and feasibility of self-report pain intensity measures for use in clinical trails in children and adolescents, *Pain*, 125, 143-157.

Stormshak, E. A. & Webster-Stratton, C. (1999) The qualitative interactions of children with conduct problems and their peers: Differential correlates with self-report measures, home behavior and school behavior problems, *Journal of Applied Developmental Psychology*, 20, 295-317.

Sturgess, J., Rodger, S. & Ozanne, A. (2002) A review of the use of self-report assessment with young children, *British Journal of Occupational Therapy*, 65, 108-116.

Teucher, U. (2008, July) Children's developing reasonings on matters of self: From transparency for parents to privileged knowledge. In P. Mitchell & U. Teucher (Chairs), *Self-knowledge and self-presentation: Individual differences and culture,* symposium conducted at the Biennial Meeting of the International Society for the Study of Behavioural Development, Wurzburg, Germany.

Tourangeau, R., Couper, M. P. & Conrad, F. (2004) Spacing, position and order: Interpretive heuristics for visual features of survey questions, *Public Opinion Quarterly,* 68, 368-393.

van Baardewijk, Y., Stegge, H., Andershed, H., Thomaes, S., Scholte, E. & Vermeiren, R. (2008) Measuring psychopathic traits in children through self-report. The development of the Youth Psychopathic Traits Inventory-Child Version, *International Journal of Law and Psychiatry*, 31, 199-209.

Van Schaik, P. & Ling, J. (2007) Design parameters for rating scales for web sites. *ACM Transactions on Computer-Human Interaction,* 14, 1, Article 4.

Weems, G. H., Onwuegbuzie, A. J., Schreiber, J. B. & Eggers, S. J. (2003) Characteristics of respondents who respond differently to positively and negatively worded items on rating scales, *Assessment & Evaluation in Higher Education,* 28, 587-607.

Yorke, M. (2009) 'Student experience' surveys: Some methodological considerations and an empirical investigation, *Assessment & Evaluation in Higher Education,* 34, 721-739.

Table 1

Overall mean scores (and standard deviations) on the scales for each of the five scenarios

	Time 1			Time 2		
	Mean	*SD*	*N*	*Mean*	*SD*	*N*
Scenario 1						
Very/highly 6 ... 1 not at all/not likely	3.69	.61	31	3.84	.81	31
Very/highly 1 ... 6 not at all/not likely	3.54	.61	33	3.53	.56	33
Not at all/not likely 6 ... 1 very/highly	3.13	.66	31	3.12	.68	31
Not at all/not likely 1 ... 6 very/highly	2.95	.69	35	3.20	.61	35
Scenario 2						
Very/highly 6 ... 1 not at all/not likely	3.62	.51	31	3.61	.62	31
Very/highly 1 ... 6 not at all/not likely	3.42	.62	33	3.43	.53	33
Not at all/not likely 6 ... 1 very/highly	3.18	.65	31	3.09	.70	31
Not at all/not likely 1 ... 6 very/highly	3.10	.62	35	3.31	.58	35
Scenario 3						
Very/highly 6 ... 1 not at all/not likely	3.77	.56	31	3.74	.68	31
Very/highly 1 ... 6 not at all/not likely	3.56	.57	33	3.42	.57	33
Not at all/not likely 6 ... 1 very/highly	3.23	.76	31	3.14	.75	31
Not at all/not likely 1 ... 6 very/highly	3.20	.73	35	3.19	.59	35
Scenario 4						
Very/highly 6 ... 1 not at all/not likely	3.51	.67	31	3.51	.58	33
Very/highly 1 ... 6 not at all/not likely	3.51	.67	33	3.63	.70	31
Not at all/not likely 6 ... 1 very/highly	3.31	.67	31	3.11	.60	31
Not at all/not likely 1 ... 6 very/highly	3.16	.66	35	3.22	.66	35
Scenario 5						
Very/highly 6 ... 1 not at all/not likely	3.58	.47	31	3.71	.59	31
Very/highly 1 ... 6 not at all/not likely	3.62	.70	33	3.35	.50	33
Not at all/not likely 6 ... 1 very/highly	3.21	.59	31	3.10	.71	31
Not at all/not likely 1 ... 6 very/highly	3.21	.51	35	3.27	.58	35

Table 2

ANOVA summary table showing the main effects for Time, scenario, anchor word, and scale and the associated error values

Source	Sum of Squares	*df*	Mean square	*F*	*p*	η^2
Time	.001	1	.001	0	.99	.00
Time * anchor word	.004	1	.004	.01	.93	.00
Time * scale	.09	1	.09	.17	.69	.00
Time * anchor word * scale	2.77	1	2.77	4.88	.03	.04
Error	71.40	126	.57			
Scenario	.50	3.30	5.50	.71	.56	.01
Scenario * anchor word	1.94	4	.48	2.74	.03	.02
Scenario * scale	.72	3.30	.22	1.02	.39	.01
Scenario * anchor word * scale	.60	3.30	.18	.85	.47	.01
Error	89.15	415.55	.22			
Time * scenario	1.16	3.74	.31	1.87	.12	.02
Time * scenario * anchor word	.38	3.74	.10	.61	.65	.01
Time * scenario * scale	.68	3.74	.18	1.09	.36	.01
Time * scenario * anchor word * scale	.31	3.74	.08	.49	.73	.00
Error	78.37	471.41	.17			
Anchor word	53.57	1	53.47	25.67	.001	.17
Scale	2.17	1	2.17	1.04	.31	.01
Anchor word * scale	3.09	1	3.09	1.48	.23	.01
Error	262.00	126	2.08			

Table 3

Confidence intervals for the scenarios and alpha values.

Scenario	Alpha value	95% Confidence Interval	
		Lower Bound	Upper Bound
Time 1			
1	.52	.40	.62
2	.48	.36	.59
3	.37	.22	.50
4	.35	.20	.49
5	.39	.25	.52
Time 2			
1	.56	.44	.66
2	.60	.49	.69
3	.59	.48	.68
4	.45	.30	.57
5	.45	.30	.57

Appendix 1 The five scenarios (adapted from Burton et al., 2003).

Note each scenario begins in the same way, but varies in the ways in which it ends.

Scenario 1

Let's pretend that one day a class was about to take a very important test. Chris and Pat sit by each other in class. The teacher gives the test to everyone and says, "Go ahead and start working on the test. I have to go to the head teacher's office to talk to someone. Please remember to keep your eyes on your own test." The teacher leaves the room and the class begins working on the test. Pat wants to do very well, but the test starts getting really hard. Pat looks over and sees Chris opening a book under the desk and cheating. Chris leans over and says, "Hey Pat this is a really hard test. Here, use my answers and you will do really well."

Scenario 2

…Pat looks over and sees Chris opening a book under the desk and cheating and hears Chris say, "Wow, this test is really easy now."

Scenario 3

…Pat looks over and sees Chris opening a book under the desk and cheating.
Chris leans over and says "Hey Pat this is a really hard test. If you do not cheat, you are going to fail."

Scenario 4

… Pat looks over and sees Chris opening a book under the desk and cheating. Chris has lots of friends and everybody likes Chris. Pat wants to be just like Chris.

Scenario 5

… Pat looks over and sees Chris opening a book under the desk and cheating. Chris turns to Pat and says, "Hey Pat, this test is a really hard test. Look at my answers. I always do great on tests."

Appendix 2

An example of one of the sets of scales used for scenario 1.

1. How much do Pat and Chris like each other?

6	5	4	3	2	1
Very Much					Not at all

2. Is Pat going to cheat like Chris?

6	5	4	3	2	1
Highly likely					Not likely

3. How upset does Pat feel?

6	5	4	3	2	1
Very upset					Not upset

4. How often does this sort of thing really happen?

6	5	4	3	2	1
All the time					Not ever

5. How upset would you be watching this happen?

6	5	4	3	2	1
Very upset					Not upset

6. If you were Pat, would you cheat?

6	5	4	3	2	1
Highly likely					Highly unlikely