Measuring Attitude with Positively Packed Self-Report Ratings: Comparison of Agreement and Frequency Scales

Gavin T. L. Brown
School of Education, University of Auckland
Running Head: Measuring Student Attitude
Contact: Private Bag 92019, Auckland 1, New Zealand
Email: gt.brown@auckland.ac.nz

Citation:
Brown, G. T. L. (2004). Measuring attitude with positively packed self-report ratings: Comparison of agreement and frequency scales. *Psychological Reports, 94,* 1015-1024.

Author Note:
The research for this article was carried out while the author was with the New Zealand Council for Educational Research and was funded through Purchase Agreement by the New Zealand Ministry of Education. I thank the Essential Skills Assessments project team at NZCER, led by Cedric Croft, for their assistance and participation in this research. Appreciation is also expressed to John Hattie, University of Auckland, for constructive criticism

Abstract

Self-report rating scales with balanced response formats, anchored with vague frequency of activity indicators, often elicit inadequate information, especially when respondents are inclined to a generally positive attitude towards the psychological object being rated. This study used an unbalanced or positively packed rating scale with both frequency and agreement response anchors within the context of a high school student (N = 734) questionnaire about studying and learning practices and conceptions. Psychometric characteristics and communication factors were investigated using 12 pairs of items for which both frequency and agreement response formats were used. Communication factors identified by Schwarz (1999), such as small changes in wording, provided adequate explanation for changes in response rate and/or fit to the IRT measurement model for three pairs of items. Although psychometric evidence of the superiority of agreement over frequency response format was not conclusive, continued use of agreement anchors with a positively packed rating scale appears justified.

The issue of response format in self-report questionnaires is not unproblematic (Gable & Wolf, 1993). Differing response formats (e.g., semantic differential, Likert agreement) have differing psychometric properties and so the question of which method is preferred is relevant (Ofir, Reddy, & Bechtel, 1987). Five to seven response points on a scale appear optimal (Comrey & Montag, 1982; McKelvie, 1978), so-called neutral mid-points should be avoided (Doyle, 1975; Ory & Wise, 1981), the end points of rating scales, at least, should be clearly labelled (Dixon, Bobo, & Stevick, 1984), and the intermediate response options should be labelled to obtain finer discriminations within a portion of the rating scale continuum (Lam & Klockars, 1982). Lam and Klockars (1982) argued that participants attended to the meaning of the intermediate response options when they found that positively packed rating scales (i.e., more response options representing the positive end of the continuum) generated lower mean scores than unlabelled or equally spaced response scales. Furthermore, the choice of adjectives to anchor response points is also critical (Gable & Wolf, 1993). The adjectives have to represent ascending values on the response format continuum. Commonly, adjectives in balanced response formats are simple mirrors of each other, but when a positively or negatively packed design is used some care is needed in selecting the intermediate adjectives.

There are clearly problems associated with how people respond to questions (Oppenheim, 1966, Sudman, Bradburn, & Schwarz, 1996), and any response that requires individuals to report how often they engage in an activity is prone to a multitude of errors. Schacter (1999) reported seven sins of memory, which are possible sources of error in honest self-reports. These included (Schacter's term is in brackets):

- we forget things over time (transience),
- we don't pay attention to everything we do (absent-mindedness),
- we sometimes can't remember things (blocking),
- we associate some things in our memory to the wrong thing (misattribution),
- our memory can be changed due to outside influences (suggestibility),
- our memory is subject to unconscious influences connected to our present beliefs (bias), and
- sometimes we can't forget memories that we'd like to forget (persistence).

In addition to these memory weaknesses, respondents may also a lack of specific episodic memory for the behaviours being questioned and instead only have fragmented recall (Schwarz, 1999). This may prevent respondents from recalling and counting how often they engage in the behaviour, resulting in a computation of a frequency estimate using inference rules.

Furthermore, how participants understand the wording of self-report questionnaires and how they interpret their own behaviour in light of questionnaire response formats also introduce error (Schwarz, 1999). As an example of the latter type of error, Schwarz (1999) identified how respondents use information gleaned from the questionnaire, for example the nature of the response scale, to shape their responses. In other words, a frequency scale for a certain behaviour that breaks response frequencies into daily, weekly, monthly, annually suggests that on average the behaviour is not very frequent leading participants to respond differently than when confronted with a response scale with frequencies such as hourly, half-daily, daily, 2-3 daily, weekly. Schwarz (1999) also argued against the use of vague quantifiers, such as 'sometimes', 'frequently', and so on since these terms can only reflect the respondents' subjective standards and not some objectively real frequency.

Thus, there is reason to believe that self-report inventories would be better served by using agreement rather than frequency response formats. Agreement formats may be less prone to memory error since they elicit information based on the present, rather than asking for recalled, knowledge, beliefs, opinions, or attitudes on the part of respondents. The purpose of this study is to add to the literature about the psychometric properties of frequency and agreement response

formats by examining both classical test theory (CTT) and item response theory (IRT) models for parallel items in differing response format conditions.

Method

This research was conducted in the context of a larger exploration of student and teacher beliefs about learning and studying (Brown, 2002a, 2002b, 2000a, 2000b). In that research a student survey of beliefs and attitudes toward various study approaches and activities was conducted. The questionnaire began with 10 open ended questions and 112 rating items. This report is restricted to 12 of the 112 rating items which were trialled with both frequency and agreement rating scales. Note since the rating items began after the initial ten items, they were numbered arbitrarily from #11. The rating items were written to fit the language and reference points appropriate to New Zealand secondary schooling. The items were reviewed by New Zealand teachers and then, pilot-tested with Year 11 students before administration.

A common rating scale was developed for the questionnaire with the goal of generating variance in the resulting data. Variance within the responses of participants is necessary in order to accurately measure any psychological object. It is likely that balanced response anchors will not provide variance when participants are inclined to respond positively to all items because they are deemed equally true or valuable. In other words, if the statement being responded to is something so socially accepted that all participants are likely to agree with it, it is difficult to elicit variance in response data because most respondents will have generally positive affect towards the psychological object. In the case of student responses to beliefs about studying as in this study, it is likely that students will view all approaches or practices as something they ought to or could agree with. For example, it is highly likely that secondary school students will tend to agree or believe they ought to agree with such actions as paying attention when studying, keeping their mind on task, memorising to learn subjects, taking notes, setting goals, sorting out confusion, etc.. Thus, a positively packed response scale (Lam & Klockars, 1982) was adopted containing four positive response points and two negative response points.

Two response formats were used to examine effect of type of response format: an agreement scale and a frequency scale. Hattie (personal communication, February, 1999) reported unpublished research (similar in method to that of Lam & Klockars, 1982) which indicated that the following adjectives would provide consistent discrimination of degrees of agreement in a positively packed rating scale (i.e., strongly agree, mostly agree, moderately agree, slightly agree, mostly disagree, and strongly disagree. The anchors for frequency were: Never or Almost Never, Rarely, Occasionally, Often, Very Often, and Always.

Students were asked to complete the questionnaire in a one-hour school period administered at their own school, assisted by their teachers (including items read aloud) using the six point positively packed rating scales. The first 52 items of the questionnaire, which probed study behaviours for which students may have had frequency knowledge based on their own experience, used a frequency format. The second 50 items, which probed beliefs or opinions about studying, used the agreement format. A subset of 12 items were asked in both response formats (Tables 1-2 for items and response formats). Note that the agreement format versions were slight re-phrasings or exact copies of study behaviour items used earlier in the questionnaire.

Responses were scored 1 to 6 with 1 indicating the first most negative score point (A) and 6 for the most positive score point (F). Thus, high scores indicate strongly positive attitudes or highly frequent behaviours. The one negatively phrased item (#71) in this part of the questionnaire was reverse scored. The data were analysed using both item response and classical test theory. ConQuest software (Wu, Adams, & Wilson, 1997) using Masters' Partial Credit Model, (Wright & Masters, 1982, p. 49) generated information on each item's fit to the underlying model. The model assumes that the single parameter of item difficulty characterises students' responses to the items and that "all items are equally discriminating" (Hambleton, Swaminathan, & Rogers, 1991, p. 13).

ConQuest generated a scale for all items, based on the log odds or logit of students agreeing with items. The weighted fit of each was indicated by a *t*-value, such that *t* values exceeding +/-2.0 are considered not to fit the underlying model at $\alpha = .05$ level. In addition to the item response theory (IRT) information, ConQuest generated classical test theory (CTT) information such as the item-total point biserial correlation and item mean and standard deviation.

Results

The questionnaire was administered to 736 Year 11 (average age was 15 years old) secondary school students in March, the second month of the New Zealand school year. After culling invalid and missing responses, there were between 696 and 716 responses per item available to analyse the effect of response format. Tables 1 and 2 show the layout, wording, and questionnaire number of the twelve pairs of items that were administered using both response formats, and the percentage of students selecting each of the six response options.

Table 1

Percentage of Responses for Questionnaire Statements by Frequency Response Format

Frequency Format Statements[a]	**Frequency Response Format**[b]					
I do this ….	*(a) Never or Almost Never*	*(b) Rarely*	*(c) Occasionally*	*(d) Often*	*(e) Very Often*	*(f) Always*
17 I attend school regularly	0.3	0.7	1.0	6.8	32.3	58.9
18 I find it easy to stick to a study schedule	9.6	15.3	22.1	27.2	17.2	8.5
19 I use a quiet place to study in	8.5	14.7	21.1	20.7	19.8	15.3
20 I complete all my work on time	1.4	5.4	17.0	26.4	35.0	14.8
29 I pay attention fully when studying	2.2	9.7	24.1	28.9	24.1	7.9
30 My mind stays on task when I do schoolwork	1.3	9.0	24.2	35.6	25.1	4.9
37 I learn new words and ideas by associating them with words and ideas I already know	3.2	12.4	31.5	28.9	15.9	8.0
39 After reading some material, I make complete notes so I can study from them later	12.7	27.3	25.4	18.8	11.2	4.7
42 I find that the only way to learn many subjects is to memorise them by heart	6.0	20.7	27.1	24.3	14.5	7.5
43 While studying I try to find answers to questions I have in mind	4.4	10.2	26.8	27.5	20.3	10.7
53 When I study I set goals for myself in order to direct my studying	8.7	19.9	28.9	23.7	13.1	5.5
59 If I get confused taking notes in class, I sort it out right after class	16.1	24.9	25.8	15.4	12.7	5.2

Notes.

a Items are displayed in style as presented to students with item numbers from questionnaire.

b Values are percentage of participants selecting each response.

Table 2

Percentage of Responses for Questionnaire Statements by Agreement Response Format

Agreement Format Statements[a]	**Agreement Response Format**[b]					
I am like this ….	(a) Strongly Disagree	(b) Usually Disagree	(c) Slightly Agree	(d) Moderately Agree	(e) Usually Agree	(f) Strongly Agree
62 I stick to a study schedule	10.4	17.6	25.9	28.3	14.6	3.3
63 I use a quiet place to study in	8.7	14.3	15.0	23.1	21	17.9
64 I complete all my work on time	2.3	5.3	13.8	19.6	38.5	20.6
65 I memorise things by heart	5.7	11.3	24.3	32.8	19.9	6.0
66 I connect new learning with things I already know	4.3	8.5	20.3	29.4	28.4	9.0
67 I answer the questions I already have in mind when I study	4.4	12.6	16.9	31.0	25.6	9.5
68 After reading something, I make complete notes so I can study from them later	10.5	22.9	22.8	24.9	14.0	4.8
69 I keep my mind on my school work	4.3	9.7	21.4	29.9	26.6	8.1
70 I go to school regularly	0.9	1.6	4.6	7.3	20.1	65.6
71 I don't pay attention fully when studying	3.7	14.7	21.8	19.6	26	9.7
72 If I get confused taking notes in class, I sort it out afterwards	9.9	17.4	19.4	23.4	20.4	9.5
73 I set goals for myself to direct my studying	9.6	15.3	22.1	27.2	17.2	8.5

Notes.

a Items are displayed in style as presented to students with item numbers from questionnaire.

b Values are percentage of participants selecting each response.

c This negatively worded item was reverse scored

Table 3 shows the relevant psychometric statistics for each items (i.e., number of respondents per item, the item mean score and standard deviation, the 1PL IRT item fit *t* value, and the CTT item-total point biserial correlation). Additionally, Table 3 shows paired item statistics; the Pearson correlation between paired items, the kappa adjusted correlation, and the factor structure with loading. Overall, there is little difference in the mean and distribution of student responses to the two different response formats. Although the Pearson correlations between matched items on average are robust (r = .83), the chance adjusted kappa statistic shows that the average correlation is somewhat tenuous (r = .25). This would indicate that the frequency and agreement response scales are measuring quite distinct psychological objects. In contrast the impact of common content can be seen in the results of an exploratory factor analysis (maximum likelihood estimation with oblimin rotation).

Table 3.

Psychometric Results for Paired Items in Agree--Frequency Response Formats

	Agreement Format				**Frequency Format**				**Pair Comparison**			
Item Numbers	*N*	*M* (*SD*)	Item Fit (*t*)	r_{pb}	*N*	*M* (*SD*)	Item Fit (*t*)	r_{pb}	Correlation (*r*)	Kappa	Factor	Loading
62 – 18	704	3.29 (1.29)	.7	.59	711	3.27 (1.24)	1.2	.48	.55	.23	IV	.61— .63
63 – 19	705	3.87 (1.55)	.6	.55	707	3.74 (1.52)	2.7	.45	.75	.33	I	1.00— .67
64 – 20	704	4.49 (1.24)	.7	.59	705	4.32 (1.16)	2.7	.43	.71	.41	III	-.80— -.87
65 – 42	699	3.68 (1.24)	.2	.59	705	3.43 (1.33)	-.1	.36	.45	.14	II	.57— .39
66 – 37	703	3.96 (1.25)	-.3	.61	709	3.66 (1.21)	.8	.47	.44	.16	II	.77— .57
67 – 43	704	3.89 (1.30)	-.9	.61	698	3.81 (1.29)	1.4	.47	.69	.17	II	.72— .65
68 – 39	702	3.23 (1.36)	1.1	.52	708	3.02 (1.36)	.5	.47	.59	.21	IV	.68— .61
69 – 30	702	3.89 (1.25)	.2	.59	714	3.89 (1.07)	1.5	.44	.49	.19	IV—III	.37— -.35
70 – 17	698	5.41 (1.01)	1.7	.32	716	5.47 (.77)	1.4	.15	.55	.50	III	-.40— -.38
71 – 29	701	3.82 (1.36)	4.5	—	711	3.90 (1.19)	1.8	.47	—	—	—	—
72 – 59	706	3.55 (1.47)	.4	.52	710	2.99 (1.42)	1.8	.48	.61	.23	IV	.43— .46
73 – 53	705	3.52 (1.41)	-.7	.60	710	3.28 (1.32)	-.3	.60	.63	.28	IV	.64— .55
All Items	654	3.88 (.80)	.68	.55	648	3.73 (.73)	1.28	.44	.59	.26		

That analysis reveals that each item loads with its partner item on the same factor with very similar loadings, except for one pair (#69—30) whose items load relatively weakly on factors IV and III respectively. One pair (#29—71: "I pay attention fully when studying" versus the reverse-scored negatively worded item "I don't pay attention when studying") has been dropped from the pair comparison because of the extremely poor fit of the negatively worded item despite reverse scoring.

The agreement format items have marginally higher mean scores and a slightly higher average item-total correlation (r_{pb} = .55) compared to r_{pb} = .44 for the frequency response items. The agreement format items had a significantly closer fit to the underlying measurement model (t = .63), in contrast to an average fit of t = 1.18 for the frequency response items. However, only three items had poor fit (i.e., t values greater than 2.0); two frequency format (#19 and 20), and the negatively worded agreement format item (#71). With this negatively worded item (#71) removed from the analysis, the average fit to the model for the eleven agreement format items is t = .34, indicating an even closer fit of the items to the underlying single dimension.

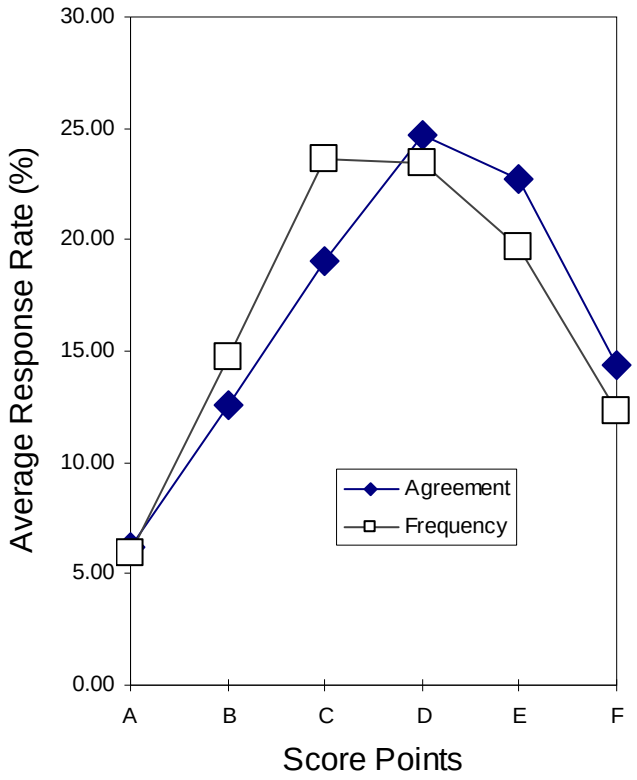

Figure 1. *Average Response Rates for All 12 Items by Item Type*

In addition to examining the psychometric characteristics of items, a close inspection of items that had changed wordings between the two formats revealed some interesting results. Item 72 (agreement) had a higher mean than parallel item 59 (frequency). The only change in wording is “right after class” (#59) became “afterwards” (#72). This suggested that the vagueness of the time marker in 72 may have had a significant impact on student response. The brevity of item 65 (agreement) relative to its parallel item 42 (frequency) may have resulted in fewer students responding negatively. This suggested the power of the word “only” in item 42 to affect students’ responses. The restriction of item 37 (frequency) to “new words and ideas” attracted fewer positive responses than the more open ended “new learning” in item 66 (agreement); a case of vaguer

wording affecting responses. Students did not respond similarly to the one item that was presented in both positive and negative wording.

The order of presentation may have played a role in the misfit of items 19 and 20 (frequency) to the model (t=2.7). These two items were the ninth and tenth of the 112 rating items presented in the questionnaire. Of the preceding eight rating items (not presented in this study) only three had fit values of $t < 2.0$. This misfit stands in stark contrast to the excellent fit of the identical items presented later in the questionnaire using the agreement response form. Perhaps a period of self-calibration was still taking place when these items were encountered. It is noted that, for both items, that the agreement format provided much more acceptable fit levels.

Conclusion

It is important to understand whether differences in student responses can be attributed to chance effects, communication factors such as the wording of statements, or to the impact of the differing response formats. Notwithstanding the possibility that the lower fit of the frequency items to the IRT model could be attributed to their early and prior presentation, there are still noticeable differences due to communication factors (i.e., wording changes) within the items studied here. Further, the interpretation of the response scale itself must still be addressed.

Three pairs of items showed the impact of communication factors identified (Schwarz, 1999) in that small changes in wording (for example, negation, introduction of restrictions or vagueness, and length of item stem) may have produced changes in response rate or fit to the measurement model. Additionally, position of presentation (i.e., early in the questionnaire and frequency before agreement) may have contributed to the poorer fit of the frequency items. Nevertheless, students were inclined to give higher ratings with the agreement response than the frequency response, despite evidence that positively packed scales generate lower mean scores than unlabelled or equally spaced response scales (Lam & Klockars, 1982). Further, the positively packed scale used here generated items with good variance in both the frequency and agreement response formats.

Overall, there were few significant differences in item psychometric characteristics (means, standard deviations, fit to IRT model) of the items responded to in two different response formats—agreement and frequency, suggesting that agreement and frequency scales generate similar responses from students. However, the very tenuous pair correlations and the content dominated factor structure suggest the response scales are not identical in effect. Additionally, researchers are left with the problem of interpreting response scales that depend on respondents remembering how often they have done some activity. This is in contrast to the clearer interpretability of agreement response scales. Thus, it is concluded that, notwithstanding similar psychometric properties, it would be better to use an agreement response scale rather than a ‘vague quantity’ frequency response scale when conducting research with self-report questionnaires. Furthermore, this study supports the use of positively packed scales in circumstances where respondents are expected to be positively oriented towards the psychological object being rated.

References

Brown, G. (2000a). *Student self reported study skills: A survey of Year 11 students.* Retrieved December 18, 2000, from New Zealand Council for Educational Research Web site: http://www.nzcer.org.nz/pdfs/7930.pdf

Brown, G.T.L. (2000b). *Year 11 Teacher Views on Student Studying.* Retrieved October 18, 2001, from New Zealand Council for Educational Research Web site: http://www.nzcer.org.nz/pdfs/8254.pdf

Brown, G.T.L. (2002a). Student Beliefs about Learning: New Zealand Students in Year 11. *Academic Exchange Quarterly, 6*(1), 110-114.

Brown, G.T.L. (2002b). *Teachers' conceptions of assessment.* Unpublished dissertation. Auckland: University of Auckland.

Comrey, A.L. & Montag, I. (1982). Comparison of factor analytic results with two choice and seven choice personality item formats. *Applied Psychological Measurement, 6*(3), 285—289.

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

Doyle, K.A. (1975). *Student evaluation of instruction.* Lexington, MA: Lexington Books, D.C. Heath.

Gable, R.K., & Wolf, M.B. (1993). *Instrument development in the affective domain: Measuring attitudes and values in corporate and school settings.* Boston, MA: Kluwer Academic Publishers.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Lam, T.C.M., & Klockars, A.J. (1982). Anchor point effects on the equivalence of questionnaire items. *Journal of Educational Measurement, 19*(4), 317—322.

McKelvie, S.G. (1978). Graphic rating scales: How many categories? *British Journal of Psychology, 69*(2), 185—202.

Ofir, C., Reddy, S. K., & Bechtel, G. G. (1987). Are semantic response scales equivalent? *Multivariate Behavioral Research, 22*(1), 21—38.

Oppenheim, A.N. (1966). *Questionnaire design and attitude measurement.* Aldershot, UK: Gower.

Ory, J.C., & Wise, S.L. (1981). Attitude change measured by scales with 4 and 5 response options. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Schacter, D.L. (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. *American Psychologist, 54*, 182-203.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*, 93-105.

Sudman, S., Bradburn, N.M., & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology*. San Francisco, CA: Jossey-Bass Publishers.

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.

Wu, M.L., Adams, R.J., & Wilson, M.R. (1997). ConQuest: Generalised item response modelling. [Software]. Melbourne: ACER.