THE GOLDEN NUMERICAL COMPARATIVE SCALE: ECONOMIES WITH PRESERVATION OF DATA QUALITY

Linda L. Golden, University of Texas at Austin
Austin, Texas 78712

KEW WORDS: Self Administered Multiobject/Multiattribute Scale Formats

1. Introduction

This research is summarizes major finding regarding the Golden Numerical Comparative Scale. The Golden Numerical Comparative Scale (GNCS) provides space and data coding economies over the bi-polar adjective scales frequently used for self administered questionnaires with no apparent loss of data quality.

Many surveys are designed to simultaneously contrast multiple objects across multiple attribute dimensions. For example, several retail stores might be compared across several image dimensions or several politicians may be compared across several issue dimensions. Implementing this multiobject/ multiattribute comparison can be cumbersome, difficult and expensive for self administered questionnaires such as mail surveys. Moreover, because of the need to repeat the scales for each object, the historically used traditional bi-polar adjective semantic differential formats (and their modified versions) can be space consuming, problematic and costly. Many researchers continue to use some version of these scales.

The major versions of the above scales for multiobject, multiattribute perceptions are variations of:

1. The *traditional semantic differential* (TSD), for which an object is rated on all attribute dimensions before the next object is rated:

Object A

Attribute 1
low _____:_____:_____high
Attribute 2
low _____:_____:_____high

2. A *modified traditional semantic differential* (MTSD), for which all objects are evaluated on a single attribute before another attribute is introduced:

Attribute 1

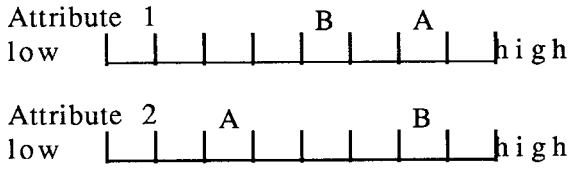
Object A
low _____:_____:_____:_____high

Object B
low _____:_____:_____:_____high

Other variants of the TSD and MTSD use categorized graphic scales or have numerical values given between the verbal anchors, however as can be seen from the above illustration, when there are multiple attributes or dimensions to consider and multiple objects to contrast along these dimensions, the TSD format and its variants can be space consuming and more expensive in terms of printing, questionnaire length, and mailing costs. Downs (1978) compared various semantic differential scales and found that respondents "preferred" the TSD scale and also found the TSD significantly "less difficult" to complete.

In order to overcome some of the deficiencies noted above for self administered multiattribute/ multiobject comparisons, Narayana (1977) posited the use of the graphic positioning scale (GPS) whereby all objects are rated on the same scale line for each pair of bi-polar adjectives. This scale also allows the respondents to provide multiple object cognizance for each of the measured attributes during the rating process. With the GPS scale, all objects are evaluated on the same scale via graphics (usually letters) reflecting their relative perceptual place-

ment on a scale between two bi-polar adjectives:



The graphic positioning scale was designed to provide a cost-effective alternative to the TSD and MTSD because of the space economy resulting from the measurement of perceptions of multiple objects on the same scale.

A potential disadvantage of the GPS is that statistical analysis using the results usually requires the analyst to transform the non-numerical graphic responses into numerical values for data analysis. This additional task can be time consuming and can provide a vehicle for introducing ambiguities and uncertainties into the analysis due to attempts to determine exactly where the letter in question is centered on the scale, and the confusion when there are several letters crowded together at a single point on the scale. There is also an increased possibility of coding errors as the questionnaires must have numbers ascertained and then recorded as opposed to just recorded.. Another disadvantage is respondents tend to become confused and response forms appear cluttered when perceptions about a number of objects on a given attribute are close together.

This paper examines a new *numerical comparative scale*, the format of which is exemplified below:

	Object A	Object B
Attribute 1:		
low 1 2 ... 7 **high**	3	2
Attribute 2:		
low 1 2 ... 7 **high**	4	6

In the GNCS format the respondents are asked to write the number that best described their impression/ attitude toward the attribute in question for each of the objects being examined. A blank is provided beside each of the attribute scales and under the name of each object for recording the response. The GNCS combines the desirable numerical properties of horizontal bi-polar adjective phrase scales with the desirable space economy and cost-efficiency of the graphic positioning scale. While data entry is generally taken for granted, this activity can be a significant consideration in estimating costs, and it is in this regard that the GPS and GNCS differ (with the GNCS easier to input). This paper presents the empirical evidence concerning other attributes of the GNCS as compared to the GPS.

2. Motivation for an Empirical Comparison of the GNCS and GPS

Mail survey is a frequently used measurement vehicle for survey analysis and the choice of a scale format can strongly influence the costs of data collection, initially through questionnaire length and subsequently through the ease of data analysis.. The format may also influence the amount of respondent effort required and the time and labor involved in preparing the data for analysis. The scale must be understandable and not so cumbersome, long, or involved as to unnecessarily increase the per respondent cost or the data compilation cost.

In order to empirically examine the relative merits of the GNCS vis a vis the other possible scale formats for self administered multiple attribute/ multiple object comparison survey questionnaires we shall compare only the GPS and the GNCS. This is done for several reasons. First of all, when the GPS has been compared with the traditional semantic differential, the GPS has been found to produce no loss of data quality (Narayana (1977); Bunder, Vincent, and Ursic (1984); Altuner, Altuner, and Chappell (n.d.). Secondly, there seems to be no evidence of reduced reliability with the GPS (Jaffe and Nebenzahl (1984); Stem and Noazin (1985). Stem and Noazin (1985) also investigated

test-retest reliability for both five and six perceptual objects on three-, five-, seven-, and nine-position TSD scales and a GPS scale and concluded that the graphic positioning scale was just as reliable as the traditional bi-polar adjective format.

Both the numerical comparative (GNCS) and GPS have an advantage over the TSD and MTSD in that perceptions of the multiple objects appear on the same scale for each dimension. This eliminates the need to reproduce the scale itself n times for k objects across l dimensions, reducing questionnaire length and potentially reducing production and mailing costs. Moreover, since the GPS has previously compared favorably with other formats for self administered questionnaires, it is only necessary to compare the GPS with the GNCS in order to provide insight into scale selection. The GPS has been compared to the GNCS in the article of Golden, Albaum and Zimmer (1987). Because the GPS has the disadvantage of requiring additional coding time and labor to convert graphic ratings to numerical ratings, if the GNCS can be shown to be produce data indistinguishable or superior to that produced using the GPS format, then the GNCS would become the preferred format for self administered multiple attribute/multiple object comparison scales.

Only Altuner, Altuner and Chappell (n.d.) have compared alternative scale formats for mail survey data, however, according to Albaum and Peterson (1985), a number of dependent measures should be used to evaluate methodological issues regarding mail survey response: response rate, cost, responses speed, data quality, and data quantity. Accordingly, the GNCS was compared to the GPS for mail survey response rate, data content, data quality, and cost considerations. Response speed was not addressed.

It was hypothesized that there would be no differences between the GPS and GNCS on the response behavior measures of response rate and item completions. Regarding response content it was hypothesized that there would be no differences in mean attribute ratings, variance of responses, and internal consistency reliability.

3. Methodology and Results

In order to compare the GPS and GNCS empirically several surveys have been performed. In the first study, a sample of 1,600 adults was selected from a nationwide consumer mail panel to represent sex, region, population density, and demographic criteria proportionate to the population of the United States. One-half of the cover letters instructed the panel member to fill out the questionnaire him/herself and the other half instructed the panel member to ask his/her spouse to fill out the questionnaire. Subjects were randomly assigned to scale-type treatments, with each treatment (GPS and GNCS) having an original sample size of 800 potential respondents.

This study focused on scaled image perceptions of three large nationwide retail store chains (Sears, K-Mart, and Montgomery Wards) and shopping frequency. For the purposes of scale comparisons discussed in this paper, however, only that part of the data dealing with multiple image comparisons across scale types will be utilized. The basis of scaled image perceptions were 19 store characteristics selected by reviewing the retail store image literature. Both scale versions (GPS and GNCS) presented the attributes in the same order, and each was scaled in seven "categories," as appropriate for the format.

Respondents receiving the numerical comparative treatment were asked to write the number from the scale that best described their perception in the blank provided for each store. Those receiving the graphic positioning scale treatment wrote the first letter of each store (S, K or W) above the point on the scale best describing their impression of the store. Questionnaire

returns were terminated six weeks after mail-out and the final sample consisted of 894 usable questionnaires.

As far as response rate was concerned, out of a total of 894 responses, more GNCS questionnaires were returned (453) than GPS questionnaires (441), however this difference was not statistically significant. For item completion rates, the pertinent rates were calculated at three distinct levels. At the most aggregate level, the proportion of respondents completing all items for each store was computed. While the completion rates illustrate that the GNCS was consistently higher than the GPS, the only significant difference occurred for Sears ($p<.01$):

All Item Completion Rates

	GNCS	GPS
Sears	81.9%	67.1%
K-Mart	65.1%	59.4%
Ward's	53.0%	49.2%

Another level of analysis for item completion rates focused on the average omission percentage per respondent for each store, as follows:

Average Omission Rate / Respondent

	GNCS	GPS
Sears	4.3%	8.9%
K-Mart	8.6%	12.8%
Ward's	32.5%	30.1%

Results of paired *t*-tests indicated that all within-scale pairs are statistically different ($p < .01$). In addition, *t*-tests for independent group comparisons yielded a significant difference in mean omission percentages between scale types for both Sears and K-Mart ($p<.05$). A final analysis of item omission rates was done for each of the 19 individual scale items and concerned the percentage of respondents omitting that item. With one exception the GNCS produced consistently lower item omission rates across scale items for both Sears and K-Mart while for Wards the reverse ordering was exhibited. Thus, in general, the GNCS dominated the GPS in terms of completion rate characteristics.

In order to assess the overall effect of scale format on image evaluation, the mean and variance of the responses to the 19 image attributes was computed for each scale. For the GNCS data these numerical responses could be entered directly from the questionnaires; however for the GPS the data had to be scored numerically for subsequent analyses. The GPS responses were coded in increments of 0.25 and transformed to the same range as the GNCS for these analyses. The data preparation was considerably more time and labor-intensive for the GPS than for the GNCS.

Fifty seven one way analysis of variance tests were run on the mean image ratings obtained using the two scale formats, and ten of these were statistically significant. This is consistent with previous research indicating that different scale formats can result in different mean ratings (Jaffe and Nebenzahl 1984), however, since the "true" mean value is unknown for the scale, it is not possible to determine if one or the other scale is a more accurate a depiction of respondent views. In this regard, to determine the relative accuracy of the different scale formats, one must turn to classical psychometric reliability theory and an examination of item response variances.

Response variance analysis and results of Bartlett's test for homogeneity of variance indicated that there were 31 scales (of a total of 57) where differences in variance between the scale formats was statistically significant at $p < .05$. As the following data show, for 30 of these 31 differences the variance of GPS was greater than that of GNCS:

	Number of Significant Scales	Number with Var(GPS) > Var(GNCS)
Sears	6	5
K-Mart	14	14
Ward's	11	11

An hypothesis postulating no significant difference in variances between scale types cannot be supported. This is important because classical psychometric reliability theory posits that the observed score Y is related to the "true" underlying unobserved score T via the equation $Y=T+\varepsilon$ where ε is the error term. (A further discussion of this model can be found in Nunnally 1967, 174-75). Presumably the variance of the true score T is the same for the two scale formats, and hence a smaller observed variance for Y using one scale format implies a smaller error variance (and hence higher reliability) for this scale format. Accordingly, the GNCS format appears in general to be more reliable (less error variance) than the GPS format.

To confirm the general tendencies observed in the scale format comparison discussed above, a second study was conducted using these two questionnaire scale formats (GNCS and GPS) in a different contextual setting. A study was developed to measure perceptions of products made in various countries (United States, Japan, Israel, East Germany and Great Britain) using 13 bi-polar attribute dimensions. The countries and attributes used in this study were the same as those used by Jaffe and Nebenzahl (1984), and the questionnaires were administered to 114 undergraduate students with respondents assigned to scale format treatments using a double changeover experimental design.

Once again the data collected from the respondents receiving the GNCS treatment could be entered into a computer directly from the survey instruments while numerical values had to be measured and assigned to the GPS data to enable analysis. For comparability with the GNCS (a one to seven range), the left-most mark on the GPS scale was scored a 1 and the right-most marked was scored a 7 and 28 intermediate sub-intervals were represented as equally spaced values between 1 and 7. (This is the most frequently used form of the GPS). An individual respondent's score was then coded by determining the sub-interval into which the graphic scale mark fell. The transformation to numerical values was coded on the questionnaires by hand and then entered into the computer data file. However, before data entry a second set of GNCS coders was utilized to checked the accuracy of the first set of coders and to make corrections as necessary. The coding processes needed to make the GPS operationally analyzable increases the costs of data preparation and potentially has data quality costs as well. Even mechanical scanners do not put the coding of the GPS on an equal footing with the GNCS coding costs.

In this second study there were 5 countries to be simultaneously examined on 13 attribute dimensions for a total of 65 comparisons possible between the GPS and the GNCS scale formats. Looking first at the means, there were only four scales, out of a total of 65, where the mean difference between the two scale formats was statistically significant at a level of .05 or less. This number is within chance levels based on the binomial distribution. There also was no pattern either to the incidence of significant difference nor to the relative directionality of the non-significant means. Thus, from the mean value perspective, the results obtained by the two formats are statistically indistinguishable.

Turning next to the comparison of variances produced by the two scale formats, (and the corresponding implications for reliability implied by the classical reliability model discussed previously), we note that with 65 comparisons there are expected to be 3.25 statistically significant differences at the .05 level and 6.77 significant differences are required before one can conclude that there is a statistically significant difference in reliability between the two scales. The results of this analysis showed that there were eight statistically significant differences, however in this study neither scale dominated the other statistically.

The GPS had statistically significant smaller error variance for five of the eight significant comparisons and GNCS had statistically significant smaller error variance for three of the eight significant comparisons. Thus, while there are measurement error differences, neither the GPS nor the GNCS emerged as statistically "better" or "worse" in this study.

4. Conclusions

This paper summarized finding comparing the GNCS to other scale formats for self administered (e.g., mailed out) questionnaires. Because previous studies have shown that the GPS offers realism and economic advantages over both the TSD and MTSD scale formats while cutting down on questionnaire length, the GPS was used as a basis for comparison with the GNCS. Both the GPS format, the GNCS provide space mailing economies over the TSD and MTSD formats for self administered questionnaires. There appears to be no loss of data quality compared with the GPS when the GNCS is used, however coding effort using GNCS was less than with the GPS with no worse reliability, and in some situations significantly better reliability than that obtained using the GPS. The GNCS also compared favorably with the GPS on response rate, data content, and data quantity measures. For self administered mail questionnaires involving multiple objects compared across multiple attributes the GNCS is superior to the TSD and MTSD scale formats and can provide an economical alternative to the GPS format by lowering data coding costs with no loss of response rate, a higher completion rate, lower item omissions per respondent, and lower measurement error (variance).

5. References

Albaum, G. and R. A. Peterson (1985), "A Paradigm for Methodological Research on Mail Survey Response," paper presented at the Annual Educators' Conference of the American Marketing Association, August.

Altuner, H. J., D. Altuner, and V. G. Chappell, Jr. (n.d.), "The Effect of Traditional versus Graphic Positioning Scales on Response Rate and Accuracy in Mail Surveys," unpublished paper.

Bunder, P., F. Vincent and M. Ursic (1984), "Graphic versus Semantic Differential Scales: An Empirical Comparison," paper presented at the annual conference of Southeast AIDS, March.

Churchill, G. G., and J. P. Peter (1984), " Research Design Effects on the Reliability of Rating Scales: A Meta-Analysis," *Journal of Marketing Research*, 21 (November), 360-375.

Downs, P. E. (1978), "Testing the Upgraded Semantic Differential," *Journal of the Market Research Society*, 20, 99-103.

Golden, Linda L. , Gerald Albaum, and Mary R. Zimmer (1987), "The Numerical Comparative Scale: An Economical Format for Retail Image Measurement," *Journal of Retailing*, 63 No. 4, 393-410.

Jaffe, F. D., and I. D. Nebenzahl (1984), "Alternative Questionnaire Formats for Country Image Studies," *Journal of Marketing Research*, 21 (November), 463-471.

Narayana, Chem L. (1977), "Graphic Positioning Scale: An Economical Instrument for Surveys," *Journal of Marketing Research*, 14 (February), 118-122.

Nunnally, Jum C. (1967), *Psychometric Theory*, New York, McGraw-Hill Book Company.

Stem, D. E., Jr., and S. Noazin (1985), "The Effects of Number of Objects and Scale Positions on Graphic Position Scale Reliability," paper presented at the annual Educators' Conference of the American Marketing Association.