A Comparison of End-Point Only Versus Fully-Verbal Labeled Response Scales

Jonathan M. Casper, North Carolina State
Myron F. Floyd, North Carolina State

Research/statistical methodology **May 30, 2009** **Poster**
Abstract 2009-182 **3:00 PM** **(Lexington/Carolina Prefunction)**

The use of rating scales with labeled endpoints is the most widely used measurement strategy in social and psychological research (Schwartz, Knauper, Hippler, et al., 1991). Likert-type rating scales also dominate opinion-based quantitative research in sport management. For example, in the 2008 issues of the Journal of Sport Management to date, seven articles reported using a five or seven point Likert-type scale. All of these articles report the use of endpoints (e.g., strongly disagree to strongly agree) for the scales, but do not use intermediate scale labels in addition to the endpoints. Several criteria exist for a Likert scale: 1) the scale contains several items; 2) response levels are arranged horizontally; 3) response levels are anchored with consecutive integers; and 4) response levels are also anchored with verbal labels which connote more-or-less evenly-spaced gradations. (Uebersax, 2006). Likert scales that indicate only endpoints may be missing several of these criteria. In particular, there are no verbal labels at all, except to define the poles of a continuum. Thus, researchers have inferred that the use of such scales be termed summated rating scales (Uebersax, 2006) which are multi-item scales that lack the full criteria of a Likert scale.

Beyond terminology, statistical research has examined how subtle differences in the presentation of a scale result in different responses. In particular, research has found similarities and differences in responses based on shading, color, and numerical label differences (Tourangeau, Couper, & Conrad, 2007), category range (Schwartz, Hippler, Deutsch, Strack, 1985); use of intensifiers (e.g., strongly agree versus agree) (O'Muircheartaigh, Gaskell, & Wright, 1993), and dependability of verbal labels (Chang, 1996). A review of literature has found no research conducted on labeling of scales by comparing scales that use endpoints only versus those that have verbal labels for each point on a scale. The purpose of this study was to compare scale types based on an endpoint-only labeled scale versus a fully-verbal labeled scale. While previous research related to scaling issues has found mixed results, it was hypothesized that there would be differences in both mean scores and variance associated with the scale types. Specifically, it was hypothesized that respondents presented with the endpoint-only scales would answer more on the poles of the scale. This tendency would lead to higher variance with endpoint-only scales but not show differences in mean values.

Data for this analysis was collected in conjunction with an undergraduate project in a statistical methods course. Students from two separate sections of the same course were asked to survey university students about barriers and interests in non-revenue varsity sports. The two scales that were compared for this study included constraints to attending non-revenue events (e.g., I don't have enough time; I don't have anybody to go with) with the endpoint-only labeled scale ranging from 1 (not true for me) to 5 (very true for me). The fully-verbal-labeled scale included 2 (usually not true for me), 3 (occasionally true for me), and 4 (usually true for me). The second scale related to interest in specific non-revenue varsity sports offered at the university. The endpoint-only scale ranged from 1 (low interest) to 5 (very high interest). The fully-verbal-labeled scale included 2 (somewhat low interest), 3 (neutral), and 4 (Somewhat high interest). One section (14 surveyors) was given questionnaires that included endpoint-only scales, while a second section (16 surveyors) was given the questionnaire with fully-verbal-labeled scales. Data collection procedures were identical between the sections. A total of 215 questionnaires were collected from the section with endpoint-only scales, while 252 were collected from the fully-verbal-labeled section. Comparison of the scales was done using three methods: 1) comparison of item and scale means based t-tests; 2) comparison of the standard deviations based Levene's test for equality of the variances; and 3) comparison of item distribution using chi-square tests.

Results showed significant differences on items for both the constraints and interest scales. For the constraints scale, three of the six item means were significantly ($p<.05$) different. Of the three items, two (I don't have enough time; I don't have anybody to go with) showed endpoint labeled scale respondents reported significantly higher means, while for the third item (I don't know about the events), the opposite was found. Two items had significantly different variances. One item (my schedule conflicts with events) had higher variance for the endpoint scale; while the other (the facilities are not nice enough) had higher variance in the labeled scale. The items for the constraints were then totaled to make an overall constraints score. No differences were found based on label type. Chi-square analysis to compare distribution differences found one item (the facilities are not nice enough) significantly different with endpoint-only respondents indicating more reposes on the poles. For the interest scale, two items (Women's Basketball, and Gymnastics) were significantly ($p<.05$) different; with endpoint labeled scale respondents reporting significantly higher means. One item (Men's Tennis) was found to have a significantly higher variance for the labeled scale. No differences were found on the overall interest means and standard deviations based on scale type. The items for the sport interest scales were then totaled to make an overall sport interest scores. Again, no differences were found based on label type. There were a total of four items (baseball, women' basketball, volleyball, and wrestling) that showed statistical distribution

variance. For two sports, endpoint-only respondents used the poles of the scale more than the verbal respondents; while for the other two, the opposite was found. Overall, nine out of the 20 items (45%) in the survey did not show any statistical differences based on labeling. A major issue with constructing survey questions is that every respondent interprets questions the same way (Dillman, 2007). While differences were found when comparing the endpoint-only and labeled scales the overall results show only subtle differences. There was no clear trend in the direction of statistical differences. For example if endpoint-only responses consistently showed higher means, standard deviations, or wider distributions, it could be concluded that scale labeling could undermine scale reliability. While subtle differences were shown, it appears that scale labeling may not be an issue.