Detecting Careless Responses to Self-Reported Questionnaires

Ronny KOUNTUR[*]

Suggested Citation:

Kountur, R. (2016). Detecting careless responses to self-reported questionnaire. *Eurasian Journal of Educational Research*, 64, 307-318 http://dx.doi.org/10.14689/ejer.2016.64.17

Abstract

Problem Statement: The use of self-report questionnaires may lead to biases such as careless responses that distort the research outcomes. Early detection of careless responses in self-report questionnaires may reduce error, but little guidance exists in the literature regarding techniques for detecting such careless or random responses in self-report questionnaires.

Purpose of the Study: The purpose of this study was to examine whether the respondent's goodness-of-fit test score (R_{GF}) can be used to indicate careless responses in completing self-report questionnaires. It is hypothesized that there is a significant difference of R_{GF} between careless responses and true responses and that R_{GF} of careless responses is higher than R_{GF} for true responses.

Method: An experimental research design that made use of a self-reported questionnaire was conducted with 205 respondents divided into two groups. The first group responded truthfully to the questionnaire while the second group responded carelessly to the questionnaire. The validity and reliability of the questionnaire had been tested. One hundred and eighty five respondents were selected as the group of true responses, while another 20 respondents comprised the group of careless responses. T-test of independent sample was used to evaluate the different R_{GF} among true responses and careless responses.

Findings: After comparing the mean scores of R_{GF} between careless responses and true responses, a significant difference was found. The

[*] Corresponding author: Dr. Ronny Kountur, Bina Nusantara University, rkountur@binus.edu

frequency distribution of true responses tends to be normally distributed while the existence of careless responses creates a skewed distribution to the right. The R_{GF} of careless responses is higher than the R_{GF} of true responses.

Conclusion and Recommendations: R_{GF} may be used as an indicator of respondent's careless responses in self-report questionnaires in which more accurate data are expected. Social science research that makes use of self-report questionnaire in measuring affective domain may compute R_{GF} to determine whether careless responses exist.

Keywords: Careless response, Questionnaire development, Random response, Goodness-of-fit.

Introduction

It is not uncommon in social science research to collect data through surveys that make use of self-report questionnaires. The use of such instruments may lead to biases (Penwarden, 2013) that distort the research outcomes (Johnson & Wislar, 2012). These biases can careless responses (Meyer et al., 2013; Summer & Hammonds, 1969; Thompson, Melancon, &Kier, 1998) and responses that may be inconsistent with the respondent's latent traits (Conijn et al., 2013). It becomes a threat to the validity of effect size estimates in correlational research (Crede, 2010) and diminishes the validity and reliability of results from survey research (Summer & Hammonds, 1969). Careless responses may be due to negative attitudes toward surveys (Rogelberg et al., 2001), sensitive items (Begin, Boivin & Bellerose, 1979; Castro, 2013), lengthy surveys (Meade & Craig, 2012), respondent gender (Sriramatr, Berry, Rodgers, & Stolp, 2012; Escobal & Benites, 2013), and poor test instructions (Rousseau & Ennis, 2013; Garcia, 2011). Early detection of careless responses in self-report questionnaires may reduce error, but as indicated by Meade & Craig (2012), little guidance exists in the literature regarding techniques for detecting such careless or random responses in self-report questionnaires.

This study introduces the use of respondent's goodness-of-fit score (R_{GF}) to detect careless responses in self-report questionnaires. This goodness-of-fit score is used to determine the consistency between the observed frequency against the expected frequency of response to items in a questionnaire. If the observed frequencies are similar to the expected frequencies, a small score will indicate that the data are consistent with a specified distribution. However, if they are sufficiently different, the score is large (Tabachnick & Fidell, 2013). In any self-reported questionnaire, the items must be reliable and internally consistent. There are two sources of inconsistency in self-reported questionnaires: the items and the respondents. When measuring the internal consistency of items in a questionnaire (making use of

Cronbach's alpha), this study assumes that respondents are consistent in their responses. If an item is inconsistent with the rest of the items, it will be removed to provide a more reliable questionnaire. After the questionnaire is tested for item consistency, it is assumed to be consistent or reliable. If - in the later use of the questionnaire - inconsistency exists, it must be caused by respondent inconsistency. This study uses goodness-of-fit scores to detecting the inconsistency of respondents answers on self-reported questionnaires while assuming that the items in the questionnaire are internally consistent. Responses that are inconsistent and that do not fit the expected and observed responses of a questionnaire may be careless responses. The individual respondent's goodness-of-fit score is derived as follows:

$$R_{GF} = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Where R_{GF} = Individual respondent's goodness-of-fit score, O_i = The value of a respondent's response for item *I*, and E_i = The expected value of item *i* that is derived from $E_i = \frac{\sum_{k=1}^{n_i} O_{i_k}}{n_i}$, where O_{i_k} = the value of response of item *i* of respondent *k*, and n_i = the total number of respondent answer item *i*.

Individual respondents who gave true responses must fit with the expected responses (in the case of a small R_{GF} value), while individual respondents who complete the questionnaire carelessly will tend not to fit with the expected responses (in which case the value of R_{GF} is larger). Note that the expected response of an item is its mean.

Some studies have been done in this area, particularly in handling the causes of careless responses such as in handling sensitive items, negative attitude toward surveys, and lengthy surveys. In handling 'sensitive items' Reaser (1975) and Begin, Boivin, and Bellerose (1979) suggested removing them from questionnaires. However, in some self-report questionnaires used in social science research, the existence of sensitive items is necessary and it is not possible to remove them. In dealing with these sensitive items, Warner (1979) devised the Random Response technique. He suggested the use of other unrelated items to estimate the answer of the sensitive items. This method of dealing with individual sensitive items produced a better estimation with a higher degree of confidentiality compared to other methods of dealing with careless responses (Begin, Boivin, & Bellerose, 1979; Crino, 1985; Lara, et al., 2006). However, there are several problems related to these two techniques (removing sensitive items and the use of the random response technique). First, this solution deals with careless responses of individual items in a questionnaire and not the careless responses of the respondent. Second, it is assumed that the researcher knows in advance which items are sensitive and which items are not sensitive, otherwise the sensitive items will remain. Third, the random response technique works well in responses that have two choices such as true-false or yes-no questions. When more responses are required, the model must be adjusted. Fourth, it is assumed that the respondent has a positive attitude towards the survey and the only reason for careless responses is due to sensitive items. In reality, some respondents have a negative attitude that leads to careless responses.

In dealing with negative attitudes toward surveys, Rogelberg et al. (2001) suggested the use of additional items that measure respondents' attitudes inserted in the questionnaire as additional items. When a negative attitude is detected, the data from this respondent may not be considered. Inserting a small number of items to detect careless responses can, in fact, detect the great majority of careless responders (Meyer et al., 2013; Meade & Craig, 2012). This technique is effective in solving the previous four problems. However, another problem exists. There is a possibility that respondents have a positive attitude toward the survey, but the respondent may get bored if the questionnaire is too long. Meade & Craig (2012) found that lengthy surveys may produce careless or random responses. For example, consider a paper/pencil Likert scale questionnaire with 160 items printed on four pages, with an equal distribution on each page. When completing such a questionnaire, the respondent may give careful responses on pages 1 and 2, but due to boredom, give random answers on pages 3 and 4. The data from that particular respondent will be considered biased. This kind of bias still cannot be identified with the recommendation given by Rogelberg et al. (2001) since the respondent had a positive attitude toward the survey but started to get bored along the way and completed it carelessly, thus distorting the result. If the respondent initially has a positive attitude, then the result of the attitude toward the survey items inserted at the beginning of the survey will be positive. This might indicate that the respondents will behave positively in completing the questionnaire, but in reality, they are not. If the attitude toward the survey items is inserted at the end, it could provide an inaccurate picture since they are based on careless responses.

Some careless responses are due to the characteristics of respondents such as those with personality disorders or addictions. The source of their careless responses is in their effortless or inattentive response behavior. Godinho, Kushnir & Cunningham (2016) suggested the use of a post-hoc detection method for screening data visually for possible error. When errors are detected, those data must be cleaned out. Unlike Keely et al. (2016), they developed an inconsistency scale that could reliably differentiate real from random responses, which may be used for those with personality disorder.

As previously mentioned, several studies have been done on careless responses, including the causes, its effect on research outcomes, and the ways of handling it. The use of negative items with equal proportion to positive items in a survey instrument to balance any careless responses has been well understood. However, it is still not possible to detect careless responses. Tatsuoka and Tatsuoka (1980) developed indices of response consistency: the norm-conformity index and the individual consistency index. These indices measure the degree of consistency between the response pattern of an individual and the difficulty ordering items in criterion referenced tests of cognitive domain but not on affective domain. No study has been done yet to identify the careless responses in completing the affective domain of a self-report questionnaire. Therefore the purpose of this study was to examine whether individual respondents' goodness-of-fit test scores (R_{GF}) can be used as an indicator to detect careless responses. It was hypothesized that there is a significant difference between R_{GF} scores of careless responses and true responses and that the R_{GF} of careless responses is higher than R_{GF} for true responses. Eliminating such respondents who give careless responses will improve the accuracy of data collected.

Method

Research Design

An experimental research design was conducted. It was a two-group post-test experimental research design. The first group got treatment of true responses while the second group got treatment of careless responses.

The groups were given a self-report questionnaire to complete. The normal time to complete the questionnaire about perception on natural medicine products was 30 minutes. The first group was given 30 minutes to complete the questionnaire with the expectation that most, if not all, would give true responses. In addition, while completing the questionnaire, they were well observed to ensure that they were truly completing the questionnaire and those who completed the questionnaire too early were rejected. The second group was given five minutes to complete the questionnaire. They were required to answer all the questions in five minutes and they were encouraged to give random answers. This was intended to give the second group a very short time period to complete the questionnaire and further encourage careless responses.

Sample

One hundred and eighty five respondents were selected using purposive sampling technique based on some criteria. The respondents were familiar with natural medicines, had been using natural medicine products, and over 18 years of age. They were considered the group of true responses since they were given proper time to complete the questionnaire (30 minutes). Another twenty respondents were asked to complete the questionnaire carelessly by giving them only five minutes to complete the questionnaire. The distribution of groups was unequal since it was assumed that those who gave careless responses were the minority. The majority of the respondents were assumed to give true responses to assure the trustworthiness of the sample.

Instrument and Procedure

The instrument was constructed as a Likert Scale. The items were developed based on theory from literature and in-depth interviews. The number of persons interviewed was based on the saturation of the information. After the seventh and eighth person, there was no more new information, thereby indicating that the information had reached saturation. Forty-five items were constructed from in-depth interview and literature reviews. These items were tested on sixty respondents for their construct validity and reliability. Items that had an Item-Reminder Coefficient (Spector, 1992) less than .30 were removed. Of 45 items, 16 were removed, leaving 29 valid items. Using Cronbach's Alpha measure of items' internal consistency, the reliability of the instrument was 0.80. The 29 items in this instrument had negative and positive statements arranged randomly. The questionnaire was given as a paper-and-pencil test. Questionnaires were collected from both groups. A mark was made on the tests of the respondents of the careless responses group to differentiate them from the true responses group.

Data Analysis

The respondent's goodness-of-fit score (R_{GF}) is derived from the formula noted in the Introduction. The higher the value of R_{GF}, the higher the possibility of having careless responses since this indicates a bad fit. The cut-off score of R_{GF} that differentiates between careless responses and true responses is determined through a line graph in the frequency distribution of R_{GF}. When there are careless responses in data, the frequency distribution of R_{GF} will be skewed to the right. When there is no careless response, the frequency distribution of R_{GF} is close to normal, as long as the number of respondents participating in the study meets the minimum requirement of central limit theorem for normal distribution. The score of R_{GF} that separates true and careless responses is the area between normal and skewed distribution. This is based on the hypothesis that the R_{GF} of careless responses is higher than the R_{GF} of true responses.

In testing this hypothesis, a *t*-test of an independent sample with a significance level of 0.05 was used. Though the sample size of the careless response group was small, 20 respondents, which may violate the normality of the distribution as required by parametric statistic, the use of parametric t-test of an independent sample is still valid. As Keller & Warrack (2016) said, "However, statisticians have shown that the mathematical process that derived the student t distribution is robust, which means that if the population is nonnormal, the results of the t test and confidence interval estimate are still valid provided that the population is not extremely nonnormal."

Results

The frequency distribution of R_{GF} was found as shown in Table 1 and Figure 1. The distribution of R_{GF} was between 1.4 and 27.7. Most of the RII frequency distribution of true responses were between 1.4 and 11.1 (73%), while few (27%) were above 11.1. All of the R_{GF} distribution of careless responses was between 09.7 and 27.7. There seems to be a different distribution between true responses and careless responses.

In an independent-sample *t*-test comparing the mean scores of RII between careless responses and true responses, a significant difference was found, $t(203)=8.06$, $p<0.05$. The mean RII of careless responses was significantly higher ($M=17.92$, $SD=3.65$) than the mean RII of true responses ($M=7.91$, $SD=5.42$). A higher RII seems to indicate careless responses.

The total frequency distribution of R_{GF} was skewed to the right. Few respondents had a higher R_{GF} score. When the frequency distribution of R_{GF} of careless response group was put in the graph together with the total frequency distribution of R_{GF}, it appears that starting where the R_{GF} of careless response exists, the tail of total frequency distribution of R_{GF} was formed as shown in Figure 2. The frequency distribution of R_{GF} values were normally distributed from a range of 0.0 to 1.4 to a range of 11.1 to 12.5 and start to form the tail of the skew from the range of 11.1 to 12.5 to the range of 26.4 to 27.7, while the frequency distribution of careless responses starts closely where the tail of the skew begins, from an R_{GF} range of 09.7 to 11.1 to a range of 26.4 to 27.7. These results show that the total frequency distribution of R_{GF} started to skew when the careless responses existed. The range of R_{GF} between 09.7 and 12.5, where the mid is 11.1, may be used as the boundary between true responses and careless responses. An R_{GF} value higher than 11.1 may be considered a careless response that should be removed from the sample.

Table 1.

Frequency Distribution of Respondent's Goodness-Of-Fit (R_{gf}) Test Score

Respondent Goodness-of-fit Score (R_{GF})	Freq. of Careless Responses	Freq. of True Responses	Total Frequency
00.0 - 01.4	0	0	0
01.4 - 02.8	0	18	18
02.8 - 04.2	0	26	26
04.2 - 05.5	0	37	37
05.5 - 06.9	0	25	25
06.9 - 08.3	0	17	17
08.3 - 09.7	0	13	13
09.7 - 11.1	1	12	13
11.1 - 12.5	2	2	4
12.5 - 13.9	0	8	8
13.9 - 15.3	1	6	7
15.3 - 16.6	2	5	7
16.6 - 18.0	2	2	4
18.0 - 19.4	5	5	10
19.4 - 20.8	5	1	6
20.8 - 22.2	0	4	4
22.2 - 23.6	1	0	1
23.6 - 25.0	0	0	0
25.0 - 26.4	1	2	3
26.4 - 27.7	0	2	2
Total	20	185	205

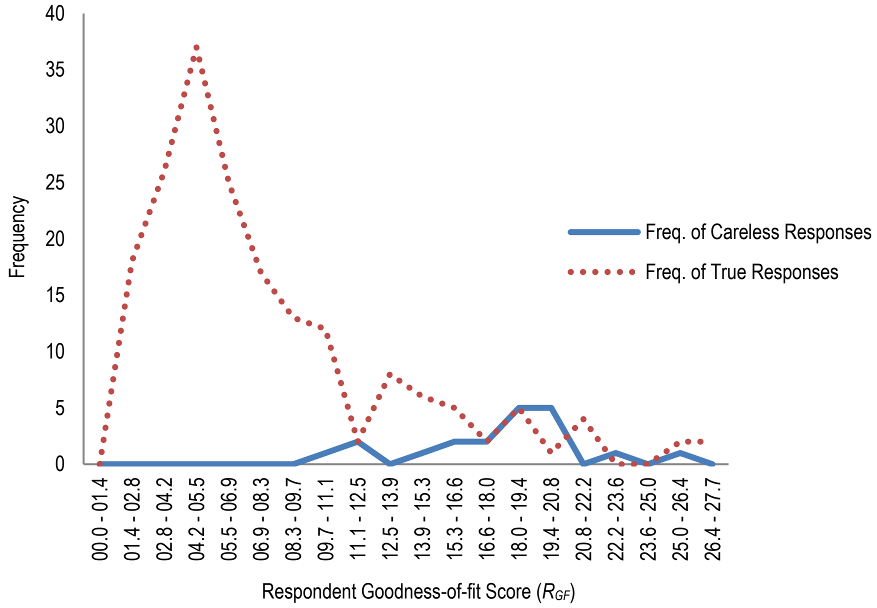

Figure 1. The Frequency Distribution of RGF Of Careless And True Responses

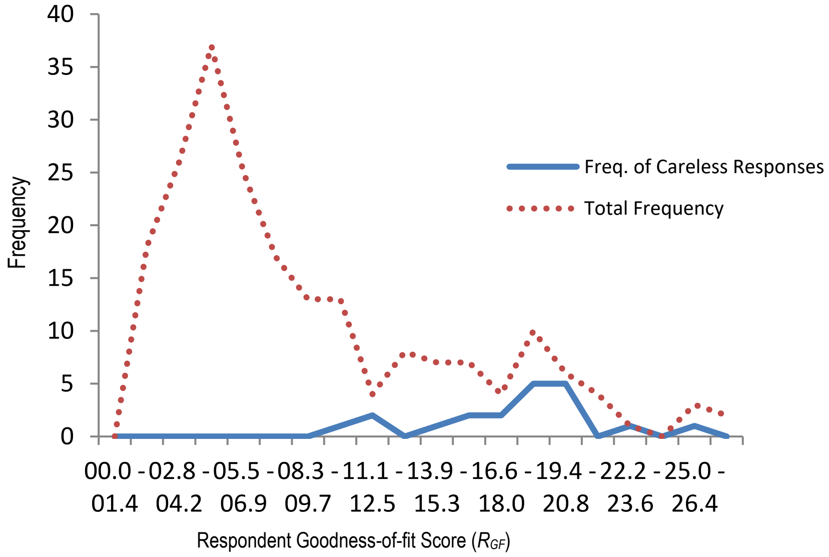

Figure 2. The Total Frequency Distribution of RGF

Discussion and Conclusion

Since the R_{GF} mean of careless responses was significantly different from the R_{GF} mean of true responses and the distribution of R_{GF} of true responses was different from the distribution of R_{GF} of careless responses, R_{GF} can be used to indicate careless responses.

Several methods of handling careless responses have been introduced, such as removing sensitive items from the questionnaire (Reaser, 1975; Begin, Boivin & Bellerose, 1979), the use of other unrelated items to estimate the answer of the sensitive items (Warner, 1979), and the use of additional items that measure respondents' attitudes toward the questionnaire (Meade & Craig, 2012; Rogelberg et al., 2001). Tatsuoka and Tatsuoka (1980) introduced the Index of Response Consistency in measuring the cognitive domain. However, no particular measurement tool has been introduced to indicate careless responses in measuring affective domain of humans in social science research. R_{GF} seems to be a promising measurement tool that may be used to indicate careless respondents in measuring the affective domain.

The data presented here indicates that the distribution of R_{GF} of true responses tends to be normally distributed, while the existence of careless responses creates a skew to the right. The data also supports the research hypothesis that the R_{GF} mean of careless responses is higher than the R_{GF} mean of true responses. Higher R_{GF} may indicate careless responses. The R_{GF} score that may be used as the starting point to indicate careless responses is the score where the R_{GF} distribution starts to skew. R_{GF} values higher than this score are considered careless responses and must be removed from the data.

The implication is that social science research that makes use of self-report questionnaires to measure the affective domain may compute the R_{GF} to determine whether careless responses exist. It is recommended to remove the answers of respondents considered to have given careless responses since such responses produce bias. These are the respondents in the distribution of R_{GF} that lay on the right tail of the skew.

This study is not without limitations. The R_{GF} is limited only to self-report questionnaires used to measure the affective domain of humans. Further study is necessary to understand self-report questionnaires used to measure other domains of human learning such as the psychomotor or cognitive domain. There is still no exact computation to determine the cut-off score of an R_{GF} that differentiates between true responses and careless responses; it makes use of visual analysis. The cut-off score is where the tail of skew started. Further study is necessary to identify a model that can determine this cut-off score.

References

Begin, G., Boivin, M. & Bellerose, J. (1979). Sensitive data collection through the random response technique: Some improvements. *Journal of Psychology*. 101(1), 53-65.

Castro, R. (2013). Inconsistent respondents and sensitive questions. *Field Methods*. 25(3), 283-298. doi: 10.1177/1525822X12466988

Conijn, J.M., Emons, W.H.M., Van Assen, M.A.L.M., Pedersen, S.S., & Sijtsma, K. (2013). Explanatory, multilevel person-fit analysis of response consistency on the Spielberger state-trait anxiety inventory. *Multivariate Behavioral Research*, 48(5), 692-718.

Crede, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, 70(4), 596-612.

Crino, M.D. (1985). The random response technique as an indicator of questionnaire item social desirability/personal sensitivity. *Educational and Psychological Measurement*, 45(3), 453-468.

Godinho, A., Kushnir, V., & Cunningham, J.A. (2016). Unfaithful findings: Identifying careless responding in addictions research. *Addiction*, 111(6), 955-956.

Keeley, J.W., Webb, C., Peterson, D., Roussin, L., & Flanagan, E.H. (2016). Development of a response inconsistency scale for the personality inventory for DSM-5. *Journal of Personality Assessment*, 98(4), 351-359.

Escobal, J., & Benites, S. (2013). PDAs in socio-economic surveys: instrument bias, surveyor bias or both? *International Journal of Social Research Methodology*, 16(1), 47-63. doi: 10.1080/13645579.2011.648420

Garcia, A.A. (2011). Cognitive interviews to test and refine questionnaires. *Public Health Nursing*, 28(5), 444-450. doi: 10.1111/j.1525-1446.2010.00938.x

Johnson, T. P., & Wislar, J. S. (2012). Response rates and nonresponse errors in surveys. *Journal of the American Medical Association*, 307(17), 1805-1806.

Keller, G. & Warrack, B. (2016). *Statistics for management and economics*, 9th ed. Australia: Thomson.

Lara, D., Garcia, S.G., Ellertson, C., Camlin, C. & Suarez, J. (2006). The measure of induced abortion levels in Mexico using random response technique. *Sociaological Method & Research*, 35(2), 279-301.

Meade, A.W., & Craig, S.B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17(3), 437-455.

Meyer, J.F., Fraust, K.A., Faust, D., Baker, A.M., & Cook, N.E. (2013). Careless and random responding on clinical and research measures in the addictions: A concerning problem and investigation of their detection. *International Journal of Mental Health and Addiction*, 7(3), 292-306.

Penwarden, R. (2013, August). How to avoid nonresponse bias. *FluidSurveys*. Retrieved from http://fluidsurveys.com/how-to-avoid-nonresponse-bias.

Reaser, J.M. (1975). A test of the forced-alternative random response questionnaire technique. *Technical Report no. 75-9*.

Rogelberg, S.G., Fisher, G.G., Maynard, D.C., Hakel, M.D., & Horvath, M. (2001). Attitudes toward surveys: Development of a measure and its relationship to respondent behavior. *Organizational Research Methods*, 4(1), 3-25.

Rousseau, B., & Ennis, J.M. (2013). Importance of correct instructions in the Tetrad test. *Journal of Sensory Studies*, 28(4), 264-269. doi: 10.1111/joss.12049

Spector, Paul E. (1992). *Summated rating scale construction: An introduction*. California: Sage Publications, Inc. p. 30-31.

Sriramatr, S., Berry, T.R., Rodgers, W., & Stolp, S. (2012). The effect of different response formats on ratings of exerciser stereotypes. *Social Behavior and Personality*, 40(10), 1655-1666.

Summers, G.F., & Hammonds, A.D. (1969). Toward a paradigm for respondent bias in survey research. *Sociological Quarterly*, 10(1), 113-121.

Tabachnick, B. G. & Fidel L. S. (2013). *Using multivariate statistics*, 6th ed. Boston: Pearson, p. 661.

Tatsuoka & Tatsuoka (as cited in Harnisch, 1981). Analysis of item response patterns: Consistency indices and their application to criterion-reference tests. *Paper presented at the Annual Meeting of the American Educational Research Association*, Long Angeles, April 13-17, 1981.

Thompson, B., Melancon, J.G., & Kier, F.J. (1998). Faking/random response scales for the PPSDQ-93 measure of Jungian personality types. *Paper presented at the annual meeting of the Southwestern Psychological Association* (new Orleans, L.A., April).

Warner, S. L. (as cited in Begin, Boivin & Bellerose, 1979). Sensitive data collection through the random response technique: Some improvements. *Journal of Psychology*, 101(1), 53-65.