Journal of Methods and Measurement in the Social Sciences,
Vol.5, No.2, 58-75, 2014

Numeric Estimation and Response Options: An Examination of the Accuracy of Numeric and Vague Quantifier Responses

Tarek Al Baghal
University of Essex

Many survey questions ask respondents to provide responses that contain quantitative information, often using either numeric open-ended responses or vague quantifier scales. Generally, survey researchers have argued against the use of vague quantifier scales. However, no study has compared accuracy between vague quantifiers and numeric open-ended responses. This study is the first to do so, using a unique data set created through an experiment. 124 participants studied word lists of paired words, where the experiment employed a 2 (context) x 2 (response form) x 6 (actual frequency) factorial design, with the context and form factors manipulated between subjects, and the frequency factor manipulated within subjects. The two conditions for the context factor are same-context and different-context conditions where the context word either was the same or different for each presentation of the target word. The other between subject factor was response form, where participants responded to a recall test using either vague quantifiers or numeric open-ended responses. Translations of vague quantifiers were obtained and used in accuracy tests. Finally, a numeracy test was administered to collect information about respondent numeracy. Different accuracy measures are estimated and analyzed. Results show context memory did not have a significant effect. Numeracy has an effect, but the direction depends on form and context. Actual frequency had a significant effect on accuracy, but did not interact with other variables. Importantly, results suggest vague quantifiers tend to improve accuracy more often relative to numeric open-ended response.

Keywords: response options, numeric cognition, vague quantifiers, recall

Surveys collect data on a wide range of information that is important for not only academic research but a range of other areas, such as marketing and policy decisions. For example, questions about television use may affect the decisions a business makes and how to invest money. Asking about how many times a person takes pain medication may affect drug policy. Importantly, many of these questions ask respondents to provide responses that contain numeric information. How these questions are asked, particularly in terms of the response format, can have an important impact on the data, including the response distribution and the overall quality of the responses. For example, the range of the scale can skew results, as people infer the average in the population. Schwarz, Hippler, Deutsch & Strack (1985) show that estimates of television watching differs on the range of the scale used. Those with low frequency scales (from up to one half-hour to more than 2.5 hours) provided a lower percentage of responses above 2.5 hours (16%) than those given a scale ranging from up to 2.5 hours to more than 4.5 hours (36%). Findings such as these are indicative that response options have an impact on the level of measurement error (i.e. variance, bias, and/or inaccuracy) in the obtained results (Biemer & Lyberg, 2003; Groves, 1989; Lessler & Kalsbeek, 1992). As such, the response format is of particular importance for ensuring that the data contains the best possible information on which to make decisions.

The research presented here uses experimental data to identify accuracy of competing response formats when asking about event frequencies. Three main

response options have been developed and used in requesting such quantitative information from respondents: numeric open-ended, numeric scales, or vague quantifier scales (Tourangeau, Rips, & Rasinski, 2000). Both numeric open-ended and numeric scale options presume that the respondent has some numeric understanding and representations of the requested information in numeric form in order to respond (Schwarz, et al., 1985). For example, asking how many hours someone watches television either using an open-ended or numeric scale such as the one discussed above presumes some knowledge of the answer in terms an amount of time (number of minutes or hours). However, numeric scales provide not only a measurement device but also an informative component as well, in that respondents infer the population average given the presented range, and may introduce an anchoring bias (Schwarz, et al., 1985).

The last response format used is vague quantifier scales. These scales use no numeric values directly, but rather verbal phrases frequently used in natural language to describe numeric data (Sanford, Moxey, & Paterson, 1994, 1996). These scales provide options that are, as the name suggests, inherently vague. For example, scale options may include words and phrases such as: “very often”, “somewhat often”, and “not very often”. As such, there is often a large variation in the numeric meaning assigned to vague quantifiers (e.g. Budescu & Wallsten, 1985). The scales also have relative meaning, such as where on the scale a respondent believes they are in comparison to similar others (Schaeffer, 1991). Based on these findings, it has been argued against using these scales when it is possible and to use numeric response options (especially numeric open-ended) instead (Beyth-Marom, 1982; Schaeffer, 1991; Tourangeau, et al., 2000).

Choice of Response Options

The response format selected for any given question should be based on some understanding of how the information is cognitively stored and represented, in order to minimize response error (Tourangeau, et al., 2000). Although arguments are made against using vague quantifier scales, there are reasons to believe that vague quantifiers are indeed better measures than numeric indicators. First, it is not clear that people are able to think numerically in a range of instances. In general, there is a lack of numeracy (numeric literacy) in the population (e.g. Galesic & Garcia-Retamero, 2010). Related to this lack of numeracy is it is cognitively more burdensome to ask about numeric information than vague quantifiers (Bradburn & Miles, 1979). Asking people for numeric responses requires the recall of all pertinent information and provision of a precise numeric value (in the open-ended formulation), which can be more difficult if understanding of numeric information is limited. Taken together, asking about exact numeric quantities likely increases the respondent’s cognitive burden substantially relative to other measures, increasing the chance of errors.

Further, research suggests that people think generally in vague quantities when thinking about numeric information. Theories such as “fuzzy-trace” and other dual-process theories suggest that people frequently rely on vague, intuitive representations of numeric information rather than on the verbatim representation of the numbers (Reyna & Brainerd, 2008). If people are thinking about information in vague quantities, asking for a precise number then requires a translation of this vague quantity into a number, which is another step that may introduce error. By asking for data in a format that is not naturally stored in memory, the response task

becomes more difficult. These theories and findings suggest that vague quantifiers may be a cognitively less demanding and more natural way of asking for numeric information.

Recent research indicates that eliciting vague quantities or general impressions may be at least as accurate as asking for numeric scale responses (Lu, Safren, Skolnik, Rogers, Coady, Hardy & Wilson, 2008). The authors obtained data on medication adherence through an electronic system that monitored the number of times the medication bottle was opened, using this opening as an indicator that the medication was taken. They asked respondents about adherence using three different scales, a six-point vague quantifier scale (from "none of the time" to "all of the time"), an eleven-point percentage scale (0, 10, 20 …100), and a six-point scale rating adherence (from "very poor" to "excellent"). The vague quantifier scale performed as well the numeric percentage scale response in relation to recorded data, with the rating scale performing better than the numeric scale, based on mean differences between reported and actual results.

The better performance of the rating scale and similar performance of the vague quantifier scale with the numeric response options may also be due to the respondents' limited numeric capabilities, but this was not testable in that research (Lu, et al., 2008). These results also conform to the finding that larger frequencies are estimated from general impressions (Conrad, Brown, & Cashman, 1998). However, it should be noted that the Lu, et al., (2008) study compared accuracy using percentage numeric scales, rather than the more standard way to ask about frequency in surveys, i.e. the number of times an activity was conducted (Tourangeau, et al., 2000). Additionally, several studies have found that respondents have problems with percentage scales in particular (Borland, 1997; Bruine de Bruin, Fischoff, Millstein, & Halpern-Felsher, 2000; Windschitl, 2002).

Therefore there is still a dearth of research on whether vague quantifiers or numeric (particularly open-ended) responses perform better in regards to accuracy, and what characteristics influence this level of accuracy. Studies examining frequencies have focused on the meanings and variations in these meanings people have placed on vague quantifiers, rather than which format performs better in predicting accuracy. To add to the extant knowledge on measurement using different response scales, an experiment was conducted in order to identify comparative accuracy of numeric open-ended and vague quantifier response options. Numeric scales are not examined given the identified biases these can create (Schwarz, et al. 1985). Further, the impact of numeracy and the contextual information of memories are examined in relation to the accuracy of the two response scales, as these may be important features in measurement of frequencies.

Method

Participants

The data comes from an experiment conducted at a large public university in the United States. The experiment is a 2 x 2 x 6 factorial design, with two between-subjects factors, each with two levels, and one within subject factor, with six levels. Subjects were selected from the university's experimental subject pool for completion of course requirements. The total number of respondents selected is based on power analyses using results from Brown (1995), which employs an experimental design on frequency estimation that is in part replicated and expanded upon in the current

research. The *F*-score from Brown (1995) testing group differences between context conditions is $F(1,38) = 14.0$, translating ($r = \sqrt{F/(F + df_D)}$) to an effect size of $r = 0.52$. Also in Brown (1995), the mean difference between rank-order correlations between estimated and actual frequency for the same and different context conditions led to the test results of $t(38) = 3.8$. This translates ($r = \sqrt{t^2/(\frac{t^2}{df})}$) to an effect size $r = 0.52$. Using these effect sizes as a guideline in a power analysis suggests at least 23 respondents per factor (Friedman, 1982). For a 2 x 2 between subjects design, as employed here, suggests at least 92 total respondents. Being somewhat conservative, 124 subjects were recruited to participate, in order to ensure adequate power if the effect sizes are not exactly equivalent to those found in Brown (1995). These subjects were randomly assigned to one of the four between-subjects combinations.

Procedure

The purpose of the experiment is to determine which factors are related to accuracy of response, in particular the effects of those with the different response formats of vague quantifiers and numeric open-ended responses. The experiment employed a 2 (context: same; different) x 2 (response form: open-ended numeric; vague scale) x 6 (frequency: 0, 2, 4, 8, 12, 16) factorial design, with the context and form factors manipulated between subjects, and the frequency factor manipulated within subjects. The within subjects factor is the number of times the target word is presented in a list, i.e. the actual frequency. Target words were presented 2, 4, 8, 12, or 16 times, for six seconds each as was done in Brown (1995).

One between-subject factor manipulated the type of context word used along with the target word. There are two conditions for this factor: same-context condition and different-context condition. This format follows the conditions used in Brown (1995) which has been shown to affect recall strategy selection and accuracy, with the different context leading to greater accuracy overall. In the same-context condition, the target word is presented with the same context word at every presentation. This context word is an exemplar for the target word. For example, for the target word CITY, it would be presented with only one context word, such as the exemplar Miami at every presentation of the word CITY. Conversely, for the different-context condition, the target word is presented in combination with a different context word at every presentation (e.g., CITY-Miami, CITY-New York, CITY-Chicago, etc.).

Target words and exemplars come from Van Overschelde, Rawson, & Dunlosky (2004) and McEvoy & Nelson (1982), studies of categories (targets) and category instances/norms (exemplars). A total of fifteen target words are used, and are selected based on similar criteria as in Brown (1995): first, the target word had to be clearly identified by a single noun, and second, each target word had to have at least sixteen category instances so that each could be presented to have any actual frequency in the different context condition.

Given the five levels of actual frequency and the fifteen words, three words are presented at each of the five actual frequencies. Word lists were created such that all fifteen target words were presented according to each of the five presentation frequencies, i.e. 2, 4, 8, 12, or 16 times. This strategy leads to 126 presentations of target words in the study list. In order to create lists where each word is presented at each level of frequency, groups of three target words (from the fifteen) were created, and these groups were varied at the five levels of frequency. This grouping leads to

five lists to ensure that each target word would be presented at each level of actual frequency (i.e. CITY presented twice on one list, four times on another, and so on through CITY being presented 16 times on one list). There were five lists for the same context condition and five lists for the different context condition. Target words were randomized in presentation throughout the lists.

The second between subject factor, and of focal interest, is the different response options offered, focusing on numeric open response and vague quantifier response options, for reasons noted above. In one condition, respondents responded to the query for the frequency of a given target word using a numeric open-ended response. The question asked, "How often did WORD appear in the presented list? _____times", where the blank was filled in by the respondent using any number.

In the vague quantifier condition, respondents were asked the same question, but instead of a blank to fill in a number, they were presented vague quantifier response options, drawn from Pohl (1981) - "Never, Not Often, Somewhat Often, Fairly Often, Quite Often, Very Often". Six vague quantifier scale points were chosen in order to match the number of actual values used. The test for both conditions included questions about the fifteen presented words as well three additional words that were not presented at all, for a test of zero presented frequencies. For this test phase of the experiment target words were asked about in a random order.

After answering all of the questions about frequency of the target words, respondents in the vague quantifier condition were asked for a numeric translation of the response options for each of the six vague quantifiers used in the above question, as suggested in Bradburn & Miles (1979). Specifically, participants were asked to translate each of the six vague quantifiers answering the question, "In the past test, how many times did you think the word WORD meant?", where WORD replaced each of the six vague quantifiers. While there is still some numeric thinking required for this method, it reduces the number of requests substantially (in this case, from 18 to 1). Further, because the question is general in nature, the task is further simplified as little or no recall is necessarily required, and the cognitive task is shifted from recall toward recognition (Tourangeau, et al., 2000).

Materials

All 124 participants studied the word lists, with word pairings of a target word and a context word. Subjects studied these words lists in groups, ranging from 5-20 respondents, presented in a classroom on a screen at the front of the room. The words were all black on a white background, presented using a timed slide show. All participants were given the instructions that word pairs would be shown, which word was to be recalled, and that a memory test for the frequency of some of the words would be given after the presentation of the list (exact instructions can be found in Appendix 1). The instructions were similar to those used in Brown's (1995) Experiment 1, in which respondents were also told that a memory test would be given after a being presented a set of word pairs. However, the instruction differed from Brown (1995) in that the nature of the memory test was not specified, whereas in this study the nature of the test was specified, largely to ensure that everyone understood expectations.

After the presentation of the word list, respondents were given a paper questionnaire, with a numeracy test and word frequency test. The numeracy test, taken from Galesic & Garcia-Retamero (2010) (available in Appendix 2), served as a filler task as well as to collect numeracy information prior to the asking about word

frequencies. Respondents were randomly assigned to different test versions, with about half receiving forms asking for vague quantifier responses and the other half asked numeric open-ended responses. Of the 124 respondents, 63 completed the numeric open-ended response form and 61 completed the vague quantifier response form. For the presentation context manipulation, 67 were presented the different context and 57 were presented the same presentation context condition.

Data Analysis

Examination of numeracy of respondents and the logical consistency of vague quantifiers and the corresponding numeric translations of these are the first results presented. Logical consistency tests examine whether respondents view vague quantifiers as implying distinctly different values and in the expected ordinal direction. However, the main outcome variable of interest of the experiment is the accuracy of the response to the question how many times a target word was presented in the list. Accuracy can be measured several different ways. Four are used in this analysis. The first two are also used in Brown (1995): the regression slope fitting estimated frequency to actual frequency and the rank-order correlations between the actual frequency and the estimated frequency. The regression slopes test the degree to which there is a bias overestimating or underestimating actual frequency. Slopes of one indicate perfect concordance between estimated and actual frequency, slopes less than one indicating underestimation, and slopes greater than one indicating overestimation. The correlation tests the relative accuracy of a participant, with larger correlations indicating higher relative accuracy, and higher (lower) estimated frequencies being related to higher (lower) actual frequencies. These two measures are estimated over all responses that a respondent gives.

Conversely, two additional important measures used here, the signed and absolute differences are calculated at the response level, with each response having a particular level of error. Therefore, the effect of actual frequency is only examinable when looking at signed and absolute differences. The signed difference identifies whether a measure is more or less error-prone in a particular direction, such as over- or underestimation of actual frequency (i.e. actual frequency – estimated frequency). Signed differences detect the direction of error, but are less easily interpreted in terms of overall error. The absolute difference better estimates overall error, and is the absolute difference between the actual and reported frequencies (i.e. |actual frequency – estimated frequency|), and is frequently used in similar recall studies (Brown, 1995; Naveh-Benjamin & Jonides, 1986).

Estimation of sample means and variances, as well as correlations and regressions on fitting estimated to actual frequency were conducted using SAS 9.2 (SAS, 2010). Estimation of hierarchical linear models using the absolute and signed difference employed HLM 7 (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011).

Results

Numeracy

Overall, the 124 respondents display a high level of numeracy. Out of the nine numeracy questions used from Galesic & Garcia-Retamero (2010), the mean number of correct responses was 7.27 (*SD* = 1.58). The median number of correct responses is eight out of nine. Thirty respondents (24.2% of the total) answered all nine correctly,

another 37 (29.8%) answered eight correctly, and 26 (21.0%) got seven correct. The minimum number correct is three, accomplished by three respondents (2.4%). There are no respondents who scored a zero in the numeracy tests. As is often done in hierarchical linear models where a variable has no zero values, numeracy was grand-mean centered for all further analyses (i.e. respondent score – 7.27) (Raudenbush, et al., 2011). This centering is done to allow better comparisons of individuals to the "average" respondent and to improve the interpretation of intercept coefficients in the hierarchical linear models.

Logical Consistency

Of the 61 respondents that answered the frequency questions using the vague quantifier scales, providing numeric translations for each of the vague terms used in the scales, 58 provided whole number translations, as intended. The remaining three respondents gave translations in terms of percentiles. Since percentiles are not the correct metric, these responses are not used in the logical consistency assessment.

The means for all six vague terms, as well as the standard error and the minimum and maximum for each are presented in Table 1 (given that the minimum values are bound at zero, the data are naturally skewed towards larger positive values). All respondents except one translated "never" as meaning zero, with this one respondent translating "never" as meaning three times. In all cases, there is complete consistency within respondents. Each respondent gave larger numeric translations along each point of the vague quantifier scale such that all translations followed the pattern: never < not often < somewhat often < fairly often < quite often < very often. This is mirrored by the means of the total sample in the below table. All mean translations are different from the others at $p < 0.01$.

Table 1
Numeric Translations for Vague Quantifier Scale

	Mean	(*SE*)	Minimum	Maximum
Never	0.05	(0.05)	0	3
Not Often	2.67	(0.16)	1	5
Somewhat Often	5.57	(0.30)	3	13
Fairly Often	8.74	(0.44)	4	15
Quite Often	12.10	(0.64)	5	25
Very Often	15.48	(0.87)	7	37

It is worth noting that the mean values in Table 1 fall remarkably close to the six possible values for the actual values of the words presented in the word lists. There is still a significant amount of variation, in part indicated by the minimum and maximum values given for each translation, with greater dispersion at the higher end of the vague quantifier scale. Even with this variation, though, the mean values of the translations for the sample are similar to the actual values of 0, 2, 4, 8, 12, and 16 used.

Accuracy

Although accuracy may be measured in a number of ways, it is first necessary to place numeric values onto the vague quantifiers that in essence conform to the distribution of the actual frequency (Lu, et al., 2008). Standard assignment of ordinal values to such scales where there is only a unit difference between scale

points, such as 1, 2, 3, 4, for a four point scale will be unsatisfactory for assessing many measures of accuracy, which is usually defined as the difference between the estimated value (the response) and the actual value (the actual frequency).

In this analysis, values are assigned to vague quantifier responses based on the individual respondents' translation of these quantifiers to numeric values (Bradburn & Miles, 1979). For example, a respondent who said that "very often" translated to 17 times, has the value 17 used for each time they selected that a word had occurred "very often", with the exception of the three cases giving percentiles noted previously. For the three respondents that gave percentile translations, all gave the translation of "never" as meaning zero and this value is used for their numeric translation for this term. For the other five vague terms, for use in assessing accuracy, the mean translated value of the total sample is imputed for these three respondents as the value of these terms. Conducting analyses with both the imputed data and the data dropping these cases show no difference between the two; therefore, only the imputed results will be presented.

Regression slopes. The first measure of accuracy, slopes of regressions of estimated on actual frequency, shows that overall respondents underestimated actual frequencies, as indicated by the overall slope mean of 0.72 (SE = 0.03). The mean does not differ significantly for those responding using vague quantifier (M = 0.71) or numeric open-ended responses (M = 0.73) as tested by a pairwise t-test $t(122)$ = 0.32, p = 0.75. Contrary to Brown's (1995) findings, there is no difference between those in the same context (M = 0.73) and different context (M = 0.71), $t(122)$ = 0.23, p = 0.81 conditions. Finally, the correlation between the slopes and numeracy (mean centered) is -0.05, and not significantly different from zero, p = 0.55.

To control for all of these effects simultaneously, and to test for possible interactions between these potentially important variables, a linear regression model is estimated, with slope of the participants as the dependent variable (not shown). The independent variables are the response form used (vague quantifier form = 1, numeric open-ended form = 0), the context words used with the target words (same context = 1, different context = 0), numeracy (mean centered), and all of the possible interactions between these three variables.[1] None of the independent variables are estimated to be significantly different from zero. Further, the omnibus F-test fails to reject the null hypothesis that the addition of any of the independent variables has an effect over the intercept.

Rank-order correlations. The results examining the rank-order correlations are nearly the same as those for the slopes. Overall, there tends to be a high correlation between estimated frequency and actual frequency, r = 0.75, suggesting an overall high level of relative accuracy. In order to test differences in correlations, correlations are first transformed to Fisher z-scores ($z = 0.5 * ln[(1 + r)/(1 - r)]$) (Fisher, 1921). Results show no difference of correlations between vague quantifier and numeric open-ended responses, $t(122)$ = 0.46, p = 0.64. There is also no difference between same and different contexts in terms of the correlations, $t(122)$ = 0.42, p = 0.68. Regressions on transformed correlations with the independent variables again being the response form, context words used, numeracy, and all of the interactions of these three (not shown) find that no significant effects for any of the independent variables at $p < 0.05$.

[1] Results were unchanged in analyses using the uncentered numeracy scores

Signed differences. To examine the effects of both the response- and respondent-level varying characteristics on individual responses, signed and absolute differences are analyzed. Examining signed differences first, Table 2 presents the mean signed difference at each level of actual frequency, i.e. 0, 2, 4, 8, 12, and 16, for each version of the response form and context condition combination that respondents fell under. For example, Vague-Same means respondents who responded using vague quantifier responses and presented the same-context condition.

Table 2
Signed Differences at Levels of Actual Frequency, by Form and Context

Actual Frequency	Vague-Same	Vague-Different	Numeric-Same	Numeric-Different
0	0.34	0.39	0.92	0.73
2	1.25	0.76	2.67	1.55
4	0.85	0.53	2.88	2.73
8	0.80	-0.94	1.36	1.50
12	-1.52	-3.26	0.05	-1.43
16	-3.80	-3.49	-3.14	-2.69
Overall	-0.34	-1.08	0.79	0.34

Across all conditions, it generally appears that at lower levels of actual frequency there are more overestimation errors, and at higher levels of actual frequency, there is greater underestimation. Overall, there is slight underestimation for vague quantifiers and slight overestimation for numeric open-ended responses; however, the overall signed differences are not significantly different from zero for the vague quantifier, same-context and numeric open-ended, different-context response combinations. Both response forms tend to a similar pattern in the directionality of errors at each level of actual frequency. However, at lower levels of actual frequency, the size of the error is smaller for vague quantifier responses, while at the two highest levels of actual frequency, numeric open-ended responses display somewhat less error.

To further examine the effect of actual frequency, as well the effect of response form, word context, and numeracy on the signed differences of each response, hierarchical linear modeling (HLM) is used (Luke, 2004; Raudenbush & Bryk, 2002). There are 18 responses for each respondent, each of which is associated with an actual presentation frequency. These responses are nested within each respondent, while numeracy, response form, and presentation context vary at the respondent level.

The two-level model has at the first level the independent variable of actual frequency. The second level of the model includes respondent effects, including numeracy (mean centered), response form, and presentation context. To ensure all interactions of the effects are accounted for, as well as main effects, both intercepts and slopes are modeled as random and used as outcomes (Luke, 2004).

The calculated intraclass correlation (ICC) of the model containing no covariates, only examining response level and respondent level variation is 0.330, suggesting that respondents account for about 33.0% of the overall variability in the observed signed differences (residual variance = 21.470). The full model including all independent variables and interactions is presented in Table 3. Robust standard errors are used (Raudenbush, et al., 2011). The inclusion of the independent

variables at both levels explains 30.3% of the residual variance (21.470-14.959/21.470 = 0.303). This reduction is moderately large, suggesting the importance these variables have in understanding signed differences of frequency report.

Table 3
Hierarchical Linear Model of Signed Differences on Response and Respondent Characteristics

Variable	Coefficient	(*SE*)
For Intercept1, β_0		
Intercept2, γ_{00}	2.22*	(0.37)
Same Context, γ_{01}	0.59	(0.63)
Form, γ_{02}	-1.15*	(0.44)
Numeracy, γ_{03}	-0.60*	(0.25)
Context*Numeracy, γ_{04}	0.88*	(0.43)
Context*Form, γ_{05}	-0.07	(0.72)
Form*Numeracy, γ_{06}	0.61*	(0.29)
Context*Form*Numeracy,γ_{07}	-0.99*	(0.47)
For Actual Frequency slope, β_1		
Actual Frequency, γ_{10}	-0.28*	(0.06)
Context*Frequency, γ_{11}	-0.01	(0.09)
Form*Frequency, γ_{12}	-0.02	(0.08)
Numeracy*Frequency, γ_{13}	0.003	(0.04)
Context*Numeracy*Frequency, γ_{14}	0.05	(0.05)
Context*Form*Frequency, γ_{15}	0.05	(0.14)
Form*Numeracy*Frequency, γ_{16}	-0.05	(0.05)
Context*Form*Numeracy*Frequency, γ_{17}	-0.05	(0.09)
Residual Variance	14.959	

n = 124, *$p < 0.05$

The results in Table 3 show that, first, the response format has an important impact on errors. Since the overall mean (intercept = 2.22) of signed errors is positive, suggesting overestimation, and the reduction caused by the main effect of vague quantifier response is not greater than the mean (thus causing the mean to become negative), the model suggests that vague quantifiers reduce error, generally bringing signed differences closer to zero, particularly at mean numeracy. Numeracy also has a significant main effect, with above average numeracy tending towards underestimation, and below average numeracy (also a negative value) tending toward errors of overestimation. However, the significant interaction between response form and mean-centered numeracy suggests that vague quantifier response tends to counteract the main effect of numeracy (if vague quantifiers are used, form = 1 and the coefficients of numeracy and the interaction sum to 0.01).

The results also show that presentation context on its own does not appear to have an effect, contrary to the findings in Brown (1995). The interaction between context and numeracy is significant, with those having above average numeracy tending towards greater overestimation in the same context condition, and those with below average numeracy tending towards underestimation. The three-way interaction between response format, numeracy, and response context is also significant. Again, this interaction suggests the effect in the use of vague quantifiers

counters the impact of the interaction of numeracy and presentation context. However, the three-way interaction is greater in size than the numeracy-context interaction, and thus slightly reverses the direction. Those with above average numeracy answering in the same presentation context using vague quantifiers slightly tend towards slight underestimation, whereas a slight tendency toward overestimation occurs among those with below average numeracy.

Greater clarity is had when examining the effect of actual frequency and the interactions between the respondent characteristics and actual frequency. First, it is evident increases in actual frequency lead to increasing likelihood of underestimation. This underestimation with increasing actual frequency is evident in Table 2 as well. However, none of the interactions with actual frequency are statistically significant. This lack of significance suggests that there is no relationship between form, context, or numeracy and actual frequency.

To greater understand the results presented in the model in Table 3, estimated mean signed differences are calculated for the four response form-context combinations at three levels of numeracy: the mean (zero), and highest (1.73) and lowest (-4.27) observed deviations from the mean. These mean signed differences are presented in Table 4. At mean numeracy, regardless of presentation context, the mean signed error is closer to zero for vague quantifiers than open-ended responses.

Table 4
Predicted Mean Signed Error by Form, Context, and Numeracy

Numeracy	Vague-Same	Vague-Different	Numeric-Same	Numeric-Different
0 (Mean)	1.59	1.08	2.81	2.22
1.73 (High)	1.41	1.09	3.29	1.18
-4.27 (Low)	2.05	1.05	1.65	4.80

When numeracy deviates from the mean of zero, the picture becomes somewhat more complicated. At high levels of numeracy, within a given presentation context (i.e. comparing response formats with same- or different-context only), vague quantifiers produce means closer to zero than numeric open-ended responses. This suggests, again, that given a context remains constant, vague quantifiers seemingly produces mean estimates with less error. However, it should be noted that open-ended responses in the different-context condition did produce mean error closer to zero than those responding to vague quantifiers in the same presentation context. Those with lower numeracy show a more complicated outcome; in the same-context condition, open-ended responses produce mean error lower than vague quantifier responses, but vague quantifier responses are less overestimated than the open-ended in the different context condition.

Subjects with lower numeracy, interestingly, show higher overestimation errors than respondents with higher numeracy in two cases, but lower overestimation errors in the other two cases, although these differences between lower and higher numeracy levels is not always substantively large (i.e. a 0.1 difference between the mean and high numeracy in the vague-different condition). Those with the lowest numeracy given different-context presentation and numeric open-ended response also had the highest overestimation error of all categorizations. However, for those with the lowest numeracy, the different-context condition and vague quantifier response led to the lowest levels of overestimation error, possibly suggesting the importance of vague quantifiers particularly among those with lower numeracy. This

possibility needs to be qualified that numeracy did not have a consistent effect on accuracy, by either response format or presentation context.

Absolute differences. Absolute error detects total error more clearly than signed differences, and is analyzed in a similar manner. First, Table 5 presents the mean absolute difference at each level of actual frequency, i.e. 0, 2, 4, 8, 12, and 16, for each version of the response form and context condition combination. Across all conditions, as actual frequency increases error also tends to increase, with all absolute differences significantly greater than zero. Overall, there is slightly less error for vague quantifiers than numeric open-ended responses, and slightly less error for the different context condition within response formats. Vague quantifiers appear to have less error at lower levels of actual frequency than numeric open-ended responses, but a slight reversal occurs at higher levels of actual frequency, with numeric responses showing somewhat smaller errors.

Table 5
Absolute Differences at Levels of Actual Frequency, by Form and Context

Actual Frequency	Vague-Same	Vague-Different	Numeric-Same	Numeric-Different
0	0.34	0.39	0.92	0.73
2	1.99	1.96	3.05	2.36
4	3.15	2.61	4.05	3.97
8	4.55	4.46	4.23	4.34
12	5.52	5.77	5.02	5.75
16	7.02	6.03	6.52	5.73
Overall	3.76	3.61	3.96	3.88

In order to examine absolute differences more completely, as was done with signed differences, a hierarchical linear model is used. The model used is identical to that used for signed differences, with the exception of using absolute differences as the dependent variable. The response- and respondent-level only model ICC is estimated as 0.151, suggesting that respondents account for about 15.1% of the variability in the observed absolute differences (residual variance = 14.870). The full model is presented in Table 6, again using robust standard errors. The inclusion of the independent variables at both levels explains 32.4% of the residual variance (14.870-10.058/14.870 = 0.324). Like the signed differences model, this reduction is moderately large, and suggests these variables are important in explaining the absolute error and accuracy of frequency reports.

Neither the main effect of response form nor presentation context is significant. The lack of impact of the context main effect is consistent across all measures of accuracy, and differs from the findings of Brown (1995). The interaction between response format and presentation context is also not statistically significant. The main effect of numeracy is significant, however, and suggests greater numeracy leads to a reduction in the size of error and lower numeracy greater error (compared to mean error, intercept = 1.69), all else being equal. This finding is consistent with the idea that numeracy is potentially an important variable in accuracy of numeric estimates. In addition, all of the interactions with numeracy are statistically significant.

Table 6
Hierarchical Linear Model of Absolute Differences on Response and Respondent Characteristics

Variable	Coefficient	(*SE*)
For Intercept1, β_0		
Intercept2, γ_{00}	1.69*	(0.41)
Same Context, γ_{01}	0.27	(0.67)
Form, γ_{02}	-0.64	(0.47)
Numeracy, γ_{03}	-0.58*	(0.27)
Context*Numeracy, γ_{04}	1.03*	(0.44)
Context*Form, γ_{05}	-0.28	(0.75)
Form*Numeracy, γ_{06}	0.58*	(0.30)
Context*Form*Numeracy,γ_{07}	-0.97*	(0.48)
For Actual Frequency slope, β_1		
Actual Frequency, γ_{10}	0.29*	(0.04)
Context*Frequency, γ_{11}	-0.01	(0.06)
Form*Frequency, γ_{12}	-0.06	(0.05)
Numeracy*Frequency, γ_{13}	0.02	(0.02)
Context*Numeracy*Frequency, γ_{14}	-0.02	(0.03)
Context*Form*Frequency, γ_{15}	0.05	(0.08)
Form*Numeracy*Frequency, γ_{16}	-0.03	(0.03)
Context*Form*Numeracy*Frequency, γ_{17}	-0.01	(0.06)
Residual Variance	10.058	

n = 124, *p < 0.05

The interaction between context and numeracy shows that the positive impact of numeracy on error only holds in the different context condition. For both the highest and lowest numeracy levels, for the same context condition, error is increased above the average. The response form-numeracy interaction, as with some of the interactions in the signed difference model, has a countering effect. In this case, the interaction counters the main effect of numeracy (coefficients sum to approximately zero). In this case, vague quantifiers bring the average error for highly numerate somewhat up, but also reduces the error of those with lower numeracy. The interaction suggests that vague quantifiers make error rates approximately equal across all levels of numeracy.

However, the significant three-way interaction between form, context, and numeracy shows that this equality depends on the context of the target to be recalled. In the different-context condition, vague quantifiers leads to expected equal error rates across numeracy, with numeracy leading to decreased error when responding in an open-ended format, via the main effect. In the same context, response by vague quantifiers reduces error for those with above average numeracy generally, but increases it for those with lower than average numeracy. The increase for those with lower than average numeracy is still offset by the context-numeracy two way interaction, such that vague quantifier response in the same-context still leads to a reduction of error from the overall mean among those with lower numeracy. The reduction in error through vague quantifier response for those with lower numeracy is just significantly greater in the different-context than the same-context.

As with the signed differences, it is evident that actual frequency affects error, with higher levels of actual frequency leading to greater error. This finding is

expected, as the greater number of events to be recalled increases amount to be recalled from memory, and this effect has been found in previous studies (Brown, 1995). However, as with the model for signed differences, none of the interactions with actual frequency is statistically significant.

Table 7 displays the effects of response form, context, and numeracy through the predicted means for the various response form-context combinations at three levels of numeracy. At the mean level of numeracy, as in signed differences, vague quantifiers reduce error levels (overall mean = 1.69) and display lower levels of error than open-ended responses regardless of context. Numeric open-ended responses lead to no change (different context) or increases (same context) in the levels of error relative to the overall mean. However, these predicted means are not significantly different, as evidenced by the lack of significance for response form, context, or the interaction between the two (as numeracy is equal to zero). Instead, it is suggestive of the possible improvements by vague quantifiers.

Table 7
Predicted Mean Absolute Error by Context, Form, and Numeracy

Numeracy	Vague-Same	Vague-Different	Numeric-Same	Numeric-Different
0 (Mean)	1.04	1.05	1.96	1.69
1.73 (High)	1.14	1.04	2.73	0.68
-4.27 (Low)	0.79	1.06	0.05	4.17

With the significant interactions including numeracy, deviations from mean numeracy (i.e. nonzero values) lead to significantly different expectations in mean error. Vague quantifiers lead to reductions in absolute error compared to the overall mean for both same and different contexts at both higher and lower levels and numeracy. The interactions lead to open-ended responses having relatively lower error rates compared to vague quantifier responses for those with higher numeracy in the different-context and for those with lower numeracy in the same context. In all other instances, however, open-ended responses either fails to reduce or increases error.

Discussion

This study is the first that has compared the accuracy between vague quantifier and numeric open-ended responses. The study also examined the impact of presentation context, respondent numeracy, and the actual frequency on error. While there appears to be no difference of response form for two measures of accuracy, overall concordance between actual and estimated frequencies (i.e. regression slopes) and the overall relative accuracy (i.e. correlations), there are impacts of response options on signed and absolute error. The findings show that vague quantifiers generally do not increase error relative to open-ended responses, and is estimated in several cases to reduce error. Vague quantifiers generally perform as well as or better than open-ended responses, with two notable exceptions.

These exceptions are that the absolute error is lower for numeric open responses when 1) context is different for highly numerate respondents and 2) when context is the same for those with lower numeracy. It is not wholly clear the reason these may occur, particularly the latter, and future research should examine the causes to any exceptions. Still, most studies will include respondents with highly varying numeracy, and ask about some memories that will have highly varying contexts, and some memories that will have very similar contexts. As such, it is unlikely that a

study will only have categorizations identified here where numeric response options are higher in accuracy. Vague quantifiers may reflect higher accuracy in recall for frequencies for a wider array of situations. These findings, in conjunction with those finding higher predictive validity for vague quantifies in frequency estimation (Al Baghal, 2014) and the expression of attitudes (Windschitl and Wells, 1996; Baghal, 2011), suggest that vague quantifiers may be a useful tool in measuring numeric quantities, contrary to previous arguments.

An additional finding of note is the influence of numeracy in accuracy of frequency recall. Although not necessarily occurring in a consistent manner, numeracy does have an impact, and should be studied further in understanding measurement of respondents' estimates. Previous studies have not examined this as a factor, but the current findings suggest incorporation of these measures. The findings here likely are in part reflective of the highly numerate sample that is used; college students in general may be expected to be more numerate, and the found numeracy scores reflect this higher numeracy. Importantly, vague quantifiers are thought to be cognitively less demanding than numeric open-ended responses (Bradburn & Miles, 1979). In this study, vague quantifiers tended to perform at least as well as numeric open-ended responses, and at times, better. If this is true among a more cognitively able sample, then for a less cognitively able sample, as might come from the general population, then vague quantifiers could perform even better than observed in this study.

Although this study suggests the importance of response form on accuracy, unlike the work of Brown (1995, 1997), altering context memory had no impact. Additional studies could examine the causes of these differences. For example, it may be that context memory may only be important as the number of items to be remembered increases, or is differentially affected by numeracy, as indicated in this study. Still, the fact that vague quantifiers performed nearly identically across memory contexts suggests these measures can be used for a number of recall tasks; however, further examination of this possibility is warranted.

The performance of vague quantifiers in this study relative to open-ended responses suggests these scales may be useful in measurement in the social sciences. Beyond frequency measurement, these findings and those showing the benefit of using vague quantifiers in subjective probability estimation (Baghal, 2011) suggest the potential efficacy in other attitude measurement. For example, verbal labeling of Likert scales may improve efficacy over simple numeric labels. The applicability should be further considered by researchers for their particular research questions. This study tested accuracy using a commonly used design, employing word lists. While many studies are not interested in memorization of word lists, the findings do suggest possible cognitive processes that may be useful in a wider array of studies. Further, if the actual number of events is of central importance, such as number of hospitalizations, then vague quantifiers would not be appropriate. Rather, these appear more useful when examining people's understandings of numbers, construction of multi-item scales, or for relationships between behaviors and other variables. Finally, it should be noted that translation techniques suggested here and by Bradburn & Miles (1979) require that the items these are applied to belong to a similar dimension. For example, "a lot" of doctor visits is likely different to "a lot" of time watching television. The choice of response formats can be based on these considerations and findings regarding the efficacy of differing options.

Author Note: Tarek Al Baghal, Institute for Social and Economic Research, University of Essex, Colchester, Essex, CO4 3SQ UK, talbag@essex.ac.uk

References

Al Baghal, T. (2014). Is vague valid?: The comparative predictive validity of vague quantifiers and numeric response options. *Survey Research Methods, 8*, 169-179.

Baghal, T. (2011). The measurement of risk perceptions: The case of smoking. *Journal of Risk Research, 14*, 351-364.

Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. *Journal of Forecasting, 1,* 257-269.

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. New York: John Wiley and Sons.

Borland, R. (1997). What do people's estimates of smoking related risk mean? *Psychology and Health*, *12*, 513-521.

Bruine de Bruin, W. D., Fischoff, B., Millstein, S. G., & Halpern-Felsher, B. L. (2000). Verbal and numerical expressions of probability: "It's a fifty–fifty chance". *Organizational Behavior and Human Decision Processes*, *81,* 115-131.

Bradburn, N. M. & Miles, C. (1979). Vague quantifiers *Public Opinion Quarterly,* 43*,* 92-101.

Brown, N. R. (1995). Estimation strategies and the judgment of event frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 1539-1553.

Brown, N. R. (1997). Context memory and the selection of frequency estimation strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 898-914.

Budescu D. V. & Wallsten T. S. (1985). Consistency in interpretation of probabilistic statements. *Organizational Behavior and Human Decision Processes*, *36*, 391-405.

Conrad, F., Brown, N. R., & Cashman, E. (1998). Strategies for estimating behavioral frequency in survey interviews. *Memory*, *6,* 339-366.

Converse, J. M. & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire*. Thousand Oaks, CA, US: Sage.

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron, 1*, 3–32.

Friedman, H. (1982). Simplified determinations of statistical power, magnitude of effect and research sample sizes. *Educational and Psychological Measurement, 42,* 521-526.

Galesic, M. & Garcia-Retamero, R. (2010). Statistical numeracy for health: A cross-cultural comparison with probabilistic national samples. *Archives Internal Medicine, 170*, 462-468.

Groves, R. (1989). *Survey errors and survey costs*. New York: John Wiley.

Lessler, J. T. & Kalsbeek, W. D. (1992). *Nonsampling errors in surveys*. New York: John Wiley and Sons.

Lu, M., Safren S. A., Skolnik, P. R., Rogers, W. H., Coady, W., Hardy, H. & Wilson, I. B. (2008). Optimal recall period and response task for self-reported HIV medication adherence. *AIDS and Behavior*, *12,* 86-94.

Luke, D. A. (2004). *Multilevel modeling*. Newbury Park, CA: Sage.

McEvoy, C. L. & Nelson, D. L. (1982). Source category name and instance norms for 106 categories of various sizes. *The American Journal of Psychology, 95*, 581-634.

Naveh-Benjamin, M. & Jonides, J. (1986). On the automaticity of frequency coding: Effects of competing task load, encoding strategy, and intention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 378-386.

Pohl, N. F. (1981). Consideration in using vague quantifiers. *The Journal of Experimental Education*. 49, 235-240.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., Congdon, R. T. & du Toit, M. (2011). *HLM 7: Hierarchical linear and nonlinear modeling*. SSI Scientific Software International: Lincolnwood, IL.

Reyna, V. F. & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, *18*, 89-107.

Sanford, A. J., Moxey, L. M. & Paterson, K. (1994). Psychological studies of quantifiers. *Journal of Semantics, 11*, 153-170.

SAS Institute, Inc. (2010). *SAS/STAT® 9.22 User's guide*. Cary, NC: SAS Institute Inc.

Sanford, A. J., Moxey, L. M. & Paterson, K. (1996). Attentional focusing with quantifiers in production and comprehension. *Memory & Cognition, 24*, 144-155.

Schaeffer, N. C. (1991). Hardly ever or constantly? Group comparisons using vague quantifiers. *Public Opinion Quarterly, 55*, 395-423.

Schuman, H. & Presser, S. (1981). *Questions and answers in attitude surveys*. New York: Academic Press.

Schwarz N., Hippler H. J., Deutsch, B. & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly, 49*, 388-395.

Tourangeau R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press.

Van Overschelde, J. P., Rawson, K. A. & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language, 50*, 289–335.

Windschitl, P. D., & Wells, G.L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied, 2,* 343-364.

Windschitl, P. D. (2002). Judging the accuracy of a likelihood judgment: The case of smoking risk. *Journal of Behavioral Decision Making, 15,* 19-35.

Appendix 1: Instructions Given to Respondents:

"This experiment is about memory, and will ask you to recall a number of words that will be presented to you on the screen at the front of the room. A pair of words will be presented. The first word is in all capital letters and is the words you will be asked to recall at the end of the experiment. The second word is to provide an example of the word you are to recall, in order to help your memory. You will be asked to recall how many times the capital words occurred in the list. If there are no questions, I will begin the word list presentation, which will take about 10 minutes. Afterward, I will give you a set of questions to answer."

Appendix 2: Numeracy Test from Galesic and Garcia-Retamero (2010)

1. Imagine that we flip a fair coin 1000 times. What is your best guess about how many times the coin will come up heads in 1000 flips?
   _____ times out of 1000

2. In the BIG BUCKS LOTTERY, the chances of winning a $10.00 prize are 1%. What is your best guess about how many people would win a $10.00 prize if 1,000 people each buy a single ticket to BIG BUCKS?
   _____ person(s) out of 1000

3. In the Daily Times Sweepstakes, the chance of winning a car is 1 in 1,000. What percent of tickets to Daily Times Sweepstakes win a car?
   _____ % of tickets

4. Imagine that we rolled a fair, six-sided die 1,000 times. Out of 1,000 rolls, how many times do you think the die would come up even (2, 4, or 6)?
   _____ times out of 1000

5. Which of the following numbers represents the biggest risk of getting a disease?
   1 in 100 1 in 1000 1 in 100

6. Which of the following represents the biggest risk of getting a disease?
   1% 10% 5%

7. If the chance of getting a disease is 10%, how many people would be expected to get the disease out of 1000?
   _____ person(s) out of 1000

8. If the chance of getting a disease is 20 out of 100, this is the same as having what percentage chance of getting the disease?
   _____ % chance

9. If person A's chance of getting a disease is 1 in 100 in 10 years and person B's risk is double that, what is B's risk?
   _____