Survey Research Methods (2014)
Vol. 8, No. 3, pp. 169-179

ISSN 1864-3361
http://www.surveymethods.org

Is Vague Valid? The Comparative Predictive Validity of Vague Quantifiers and Numeric Response Options

Tarek Al Baghal
Institute of Social and Economic Research, University of Essex

A number of surveys, including many student surveys, rely on vague quantifiers to measure behaviors important in evaluation. The ability of vague quantifiers to provide valid information, particularly compared to other measures of behaviors, has been questioned within both survey research generally and educational research specifically. Still, there is a dearth of research on whether vague quantifiers or numeric responses perform better in regards to validity. This study examines measurement properties of frequency estimation questions through the assessment of predictive validity, which has been shown to indicate performance of competing question formats. Data from the National Survey of Student Engagement (NSSE), a preeminent survey of university students, is analyzed in which two psychometrically tested benchmark scales, active and collaborative learning and student-faculty interaction, are measured through both vague quantifier and numeric responses. Predictive validity is assessed through correlations and regression models relating both vague and numeric scales to grades in school and two education experience satisfaction measures. Results support the view that the predictive validity is higher for vague quantifier scales, and hence better measurement properties, compared to numeric responses. These results are discussed in light of other findings on measurement properties of vague quantifiers and numeric responses, suggesting that vague quantifiers may be a useful measurement tool for behavioral data, particularly when it is the relationship between variables that are of interest.

Keywords: vague quantifiers; numeric responses; behavioral frequencies; predictive validity

1 Introduction

When asking about quantitative information in surveys, such as academic or other types of behaviors, there are a number of ways to provide response options to respondents. Three main response options have been developed and used in requesting quantitative information from respondents (although others exist): numeric open-ended, numeric scales, and vague quantifier scales (Tourangeau, Rips, & Rasinski, 2000). The open-ended approach leaves the response formulation and reporting to the respondent, whereby, for numeric questions, the respondent generally responds with one number. For scale responses, respondents choose a scale point most closely associated with their formulated answer. For objective measures, such as frequencies, the response options can include scale points for distinct, individual values, (e. g. 1 time) or for ranges of values (e. g. 1 to 5 times). Both 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, Hippler, Deutsch, & Strack, 1985). However, some have argued against the use of numeric scales as it may bias respondent answers as the scale provides not only a measurement device but also an informative component as well (Schwarz et al., 1985).

The last response format frequently used is vague quantifier scales. Vague quantifier scales are argued against by many survey researchers, who suggest instead that numeric open-ended responses be used (Beyth-Marom, 1982; Schaeffer, 1991; Tourangeau et al., 2000). They suggest avoiding vague quantifiers for several reasons. First, these scales provide response options that are, as the name suggests, inherently vague. Due to this, there is often sizable variation in the numeric translation assigned to vague quantifiers generally (e. g. Bradburn & Miles, 1979; Budescu & Wallsten, 1985; Schaeffer, 1991), which has also been found in education behavioral data (Cole & Korkmaz, 2013; Pace & Friedlander, 1982). Further, the scales have relative meaning, such as where on the scale a respondent believes they are in comparison to other similar individuals (Schaeffer, 1991). Additionally, vague quantifiers have different meanings for different targets (Windschitl & Wells, 1996). For example, "a lot" of risk from smoking may be different from "a lot" of risk from arsenic. These arguments against the use of vague quantifiers have also been used to question the validity of a large,

Tarek Al Baghal, Institute for Social and Economic Research, University of Essex, Colchester (Essex), CO4 3SQ (talbag@essex.ac.uk)

national student satisfaction survey, the National Survey of Student Engagement (NSSE) (Porter, 2011).

Although many argue against the use of vague quantifiers, there are also reasons that these scales may be preferable to numeric open-ended responses. First, it is not clear that most respondents are able to think quantitatively and apply numerical concepts. Studies show that a sizable portion of the overall population lack numeracy (numeric literacy) (e. g. Galesic & Garcia-Retamero, 2010). This lack of numeracy can affect the ability to respond to questions about numeric quantities accurately (Galesic, Garcia-Retamero, & Gigerenzer, 2009). Second, 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 verbatim representation of the numbers (Reyna & Brainerd, 2008). Third, additional research finds that the relationship between subjective beliefs and behaviors is stronger when using vague quantifier scales than numeric responses (Al Baghal, 2011; Windschitl & Wells, 1996). Fourth, although there is variation in interpretation of vague quantifiers, research on student surveys (including the data used in the current research) shows there is logical consistency in interpretation in meaning of vague quantifiers (Cole & Korkmaz, 2013; Nelson Laird, Korkmaz, & Chen, 2008).

Finally, and of particular importance, it is argued that it is more cognitively burdensome to ask about numeric information than vague quantifiers (Bradburn & Miles, 1979). Although respondents may not have a clear definition of what a vague term means, they do likely comprehend what the words mean in regards to what the question is asking them to report, and the other components of the survey response process may benefit by the use of vague quantifiers. Vague quantifiers do not place as much burden on the recall phase, as exact enumeration is not required (Burton & Blair, 1991; Tourangeau et al., 2000). Since recall is minimized compared to enumeration, less burden is likely to occur at the judgment stage as well (Tourangeau et al., 2000). The response selection portion of the process requires either selecting a choice among vague quantifiers or providing a number, and cognitive burden may not be different. Social desirability problems that are pointed out in student surveys (Porter, 2011) would likely be the same for either vague quantifiers or open-ended answers, as higher (lower) on the scale or higher (lower) numbers would be better for positive (negative) behaviors (Tourangeau et al., 2000).

Even though there are arguments that have been put forth for and (mainly) against the use of vague quantifiers, few studies have compared the validity of numeric open-ended and vague quantifier responses. Although not comparative, since numeric responses were not available, Carini, Kuh, and Klein (2006) shows that the vague quantifier scales contained in the NSSE do have predictive validity in relation to important outcomes, such as grades. The one study identified comparing predictive validity of vague quantifiers and numeric responses in a representative national survey did so for subjective (attitudinal) measures, finding that vague quantifiers display higher levels of predictive validity (Al Baghal, 2011). Lu et al. (2008) compare the accuracy of vague quantifiers and scaled numeric estimates for behavioral data, finding no difference between the measures. However, this study uses an 11-point percentage scale (0-100 by tens), rather than the actual number of times an activity was conducted, which may be more commonly used, given the noted problems with scales. Conversely, recent experimental data using more standard numeric measures suggest that vague quantifiers may be more accurate generally in frequency estimation (Al Baghal, 2014).

Overall, the evidence on the comparative strengths of vague quantitative and numeric open-ended responses is limited. The current research furthers the understanding of these competing measures, in an area that has yet to be explored for these measures in behavioral frequency estimation. Specifically, it examines which response format, if either, is stronger in terms of predictive validity, i.e. the relationship of the target response to other theoretically related variables. In many instances it is the relationships between variables that may be of interest rather than the simple univariate statistics, and predictive validity tests have been used to examine measurement properties of competing question formats (Chang & Krosnick, 2003; Diefenbach, Weinstein, & O'Reilly, 1993; Weinstein & Diefenbach, 1997). To date, no studies have been identified comparing vague quantifiers and numeric open-ended responses in regards to predictive validity for behavioral frequency estimates; this research does so using the NSSE, a widely used student survey, which has come into question in part due to its use of vague quantifiers.

2 Data and Methods

The data comes from the National Survey of Student Engagement (NSSE). The NSSE is an annual survey that collects data from college students in the United States from hundreds of participating institutions, with several hundred thousand surveys collected, including 335,000 collected in 2013 (National Survey of Student Engagement, 2013). The survey collects data from randomly selected college freshmen and seniors at the participating institutions. Although freshman and seniors are the target population, some respondents come from other classes. The primary purpose of the survey is twofold: 1) to assess the time and effort undergraduate degree-seeking students spend on educational activities, and 2) to assess what schools are doing to focus student efforts to these activities (National Survey of Student Engagement, 2013). The NSSE measures have been used to develop five scales intended to gauge student and institutional performance internally and comparatively (Kuh, 2003). These scales (or the constituent measures) have been used to predict important outcomes such as critical thinking and measures of academic achievement (such as grades) (Carini et al., 2006; Gellin, 2003; LaNasa, Olson, & Alleman, 2007). Data from these surveys are collected via one of two survey modes, either by a paper or web survey.

The particular NSSE data used in this research comes from 2006, when additional responses were collected on questions using vague quantifiers. The data comes from the available student respondents to the web NSSE survey where the experiment was conducted, which represent 26,204 first-year and 36,263 senior students who were randomly selected from 149 institutions (Nelson Laird et al., 2008). The NSSE is proprietary, and the data obtained for this research is a random subsample of 10,767 web respondents, due to licensing restrictions.

Some schools participating in the web survey also sent non-respondents a paper version as a follow-up. Those responding to the paper version are not included in the experiment and not analyzed in this research. Response rates are calculated by the NSSE, as the complete sample frame is not available as part of the licensing restrictions, although questionnaires, the explanation of the design, response rates, and the method used to calculate these response rates are freely available.[1] Response rates are calculated first for each participating institution, calculated as the number of respondents divided by adjusted sample size. The sample is adjusted for non-deliverable mailing addresses, students for whom contact information was not available and other students who were sampled yet unavailable during the survey administration. The final response rates reported are the average response rates across all institutions. The response rate for where only a web survey was offered in the 2006 survey was 41%, while it was 39% in schools where a paper follow-up was also offered (National Survey of Student Engagement, 2006).[2]

Analyses of respondents' answers focus on the predictive validity of the different responses in regards to variables relating to the academic outcomes and perception of educational experience. The twelve questions available for these analyses belong to two of the benchmarks, one for active and collaborative learning and one for student-faculty interaction Carini et al. (2006), Kuh (2003), Nelson Laird et al. (2008). Of the twelve items, seven are for the active-collaborative learning scale with the remaining five belonging to the student-faculty interaction scale (see Appendix A, B). The questions were first asked using vague quantifier response options (see Appendix D for an example). Then at the end of the web survey (these questions were repeated at the end of the web survey only) students were asked to quantify their response for each question by filling in a number into an open-ended response space to indicate the number of times intended by the vague term and asked to select the rate they intended (e. g. daily, weekly, monthly). For example,

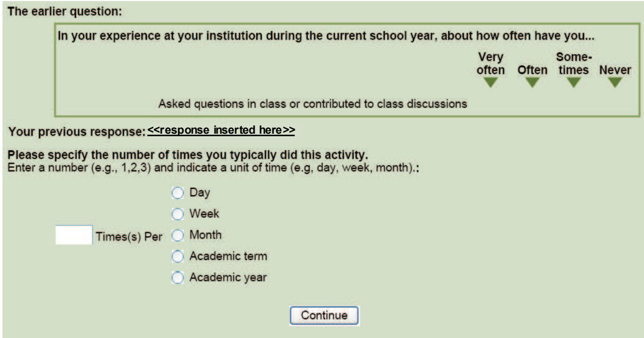


Figure 1. Example question for absolute frequency intended by vague quantifier

a person saying an event occurred "very often" was asked how many times this event occurred and if this number occurred either per day, week, month, academic term, or academic year. Thus, students entered a numeric response and selected an appropriate time frame. Figure 1 displays an example from the web survey showing the way this question was asked to respondents (from Nelson Laird et al., 2008).

Given the differences in the rate of occurrence selected by the respondent in the numeric translation of the vague quantifier response, that is, whether the number of times occurred per day, week, month, academic term or academic year (see Figure 1), it is necessary to transform numeric open-ended responses to a constant time frame, such as is done in Nelson Laird et al. (2008). In this case, all numeric answers were transformed to be on a per week time frame, through either multiplying answers by five (for those that said number per day) or dividing by the appropriate number for those saying per month, academic term or year. Specifically, time frames were adjusted by using the following multipliers: day = 5, week = 1, month = 0.231, academic term = 0.067, and academic year = 0.033. The twelve questions asked about in this manner are presented in Appendix A and Appendix B.

As the numeric responses were only obtained in the internet survey, only web respondents are used in the analyses. After transforming these respondents' numeric data to weekly rates, the distributions of these numeric translations were examined. Visual inspection showed that some responses are extreme and not plausible (e. g. an event occurring 50,000 times a week). Overall, these extreme responses are few. The data were cut at the 99th percentile of the distribution, which would in all cases lead to more reasonable responses with a minimum of data cut. For nine of the numeric translations, the use of the 99th percentile leads to cuts of translations greater than 10. For frequency of questions

[1] See http://nsse.iub.edu/

[2] More information on calculation of response rates can be found at http://nsse.iub.edu/pdf/2006_Institutional_Report/NSSE2006Overview.pdf

asked in class, the 99^{th} percentile is 50 times per week. It is 15 time per week for working in class with other students. For discussed ideas outside of class with others, the 99^{th} percentile is 20 times per week. These cut data are used in all following analyses. Further, in order that the vague quantifier scales be anchored at zero, all of the vague quantifier responses are scaled from zero ("never") to three ("very often").

For both response formats, the responses to each question are summated to create the scale value for each respondent, as the responses for both the vague quantifier and numeric response translation response options have both been used for the scales Nelson Laird et al. (2008). Since the numeric responses scales are not normally distributed, these data are standardized by first calculating z-scores (by taking the difference from the overall mean and dividing by the standard deviation). The z-scores for each question were then summed up to form the numeric scale, and given the non-normality, the logarithm was taken (adding the minimum value and a small value, i.e. 0.001 to overcome negatives and zeros).[3] Scales are summated for the vague quantifier questions by placing values on each of the response options, from 0 through 3. This assignment was used for two reasons; first, because others have summed the scale in this way i.e., Carini et al. (2006) and second, because numeric translation is needed for accuracy assessments but is not needed for predictive validity assessments.

Responses were then summed up based on the questions for each of the two scales for each respondent. Since both reports to the same activity are given by each respondent, there is a certain lack of independence in responses. However, a similar method of asking the same respondents the same outcomes using different response options is employed elsewhere in question format research (Al Baghal, 2011; Diefenbach et al., 1993), with an additional study suggesting this type of design does not impact results (Weinstein & Diefenbach, 1997). Similarly, multitrait-multimethod (MTMM) studies often ask the same respondents the same (or similar) questions multiple times with different response scales in order to asses question validity (e. g. Revilla & Saris, 2012).

The three outcome measures of interest are grades and two satisfaction measures. These are selected as it has been argued that the NSSE benchmarks are intended to be used to assess institutional performance (Kuh, 2003). Grades are used by other research to examine predictive validity of NSSE scales (Carini et al., 2006). In the current data set, only self-reported grades are available (measured on an 8-point scale, C- or lower to A). However, self-reporting grades may induce social desirability biases (Porter, 2011). As such, two additional measures are also used; satisfaction with the educational experience has been noted to be as a potentially important educational outcome and construct, are indicated in the NSSE by two measures (Carini et al., 2006). The first is with overall college experience and the second being if the student would choose to attend the institution again if they could "start over", both measured on four-point scales (Poor-Excellent, Definitely No-Definitely Yes, respectively) (see Appendix C for exact wordings). Unlike grades, which are reflective of the student's performance, possibly inducing social desirability biases, satisfaction is more likely reflective of the institution's performance, and hence potentially less prone to social desirability issues (Tourangeau et al., 2000). The goal is that by using three potentially important variables, consistency in the findings will provide stronger evidence than only one would.

It is important to note that the data from the NSSE may be similar to a cluster sample for surveys (Kish, 1965). This clustering could arise from the fact that respondents are selected within participating institutions (the clusters). Respondents are selected and grouped within the 127 universities included in the survey, and the design effect (*deff*) is estimated for each of the means of the variables of interests. This includes the 12 vague quantifier responses to the two scales, the 12 numeric translations for each of these responses, and the two related variables of interest, grades and satisfaction, for a total of 27 design effects estimated. Design effects are estimated based on the Taylor series approximation variance estimates for the clustered survey design (Kish, 1965).

The results show that, overall, clustering is an issue that must be accounted for in the NSSE. The calculated design effects range from 1.104 to 9.267. Only three of the 26 estimated *deffs* are less than 2, with the mean *deff* being 3.888 (median = 3.518). This mean *deff* suggests, on average, a near four times increase in the estimated variance due to the clustered design compared to the simple random sampling assumptions frequently used in analyses. Given the evident clustering effects on the variance estimation, it would be inappropriate to use simple random sampling assumptions in variance estimation, and hence hypothesis testing, for the remaining analyses. Therefore, appropriate estimation procedures are employed using the SAS system (SAS Institute, Inc., 2010). Variances will all be estimated using the Taylor series approximation.

3 Results

Predictive validity has been shown to be an important aspect of the measurement properties of frequency questions in past research (Chang & Krosnick, 2003; Lu et al., 2008). This type of validity can be measured via the relationship between the measures of interest and theoretically related variables. In this case, the measures of interest are the active-

[3]For the numeric translation scale, a simple addition of the numeric responses given was also examined in forming the scales. The results in terms of directionality and significance were near identical to the z-score transformed scales.

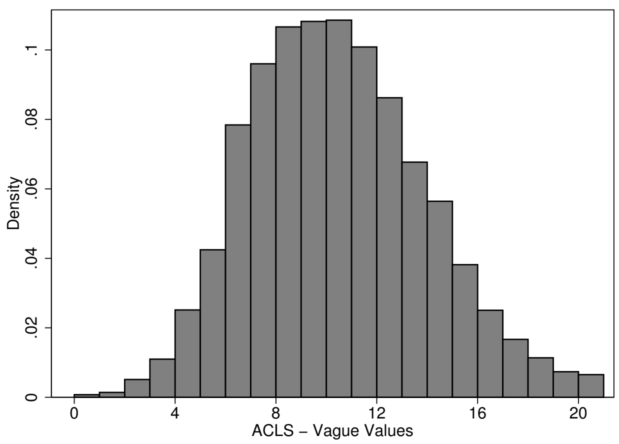


Figure 2. Distribution of ACLS-Vague Scale

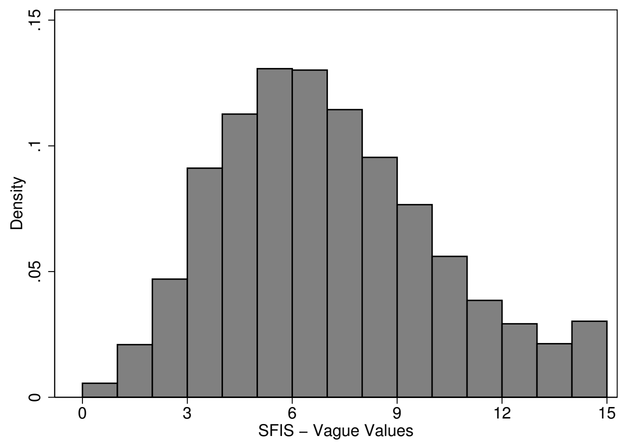


Figure 4. Distribution of SFIS-Vague Scale

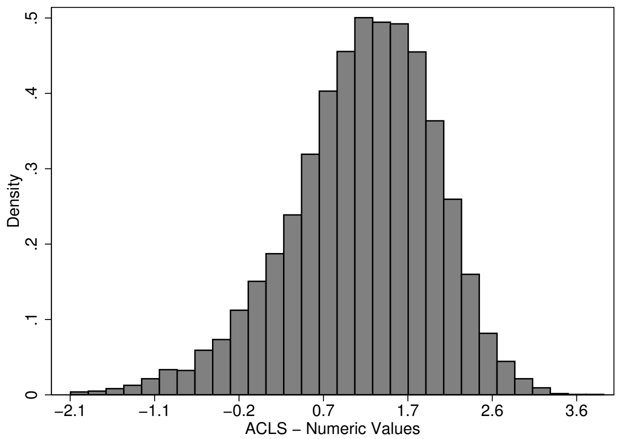


Figure 3. Distribution of ACLS-Numeric Scale

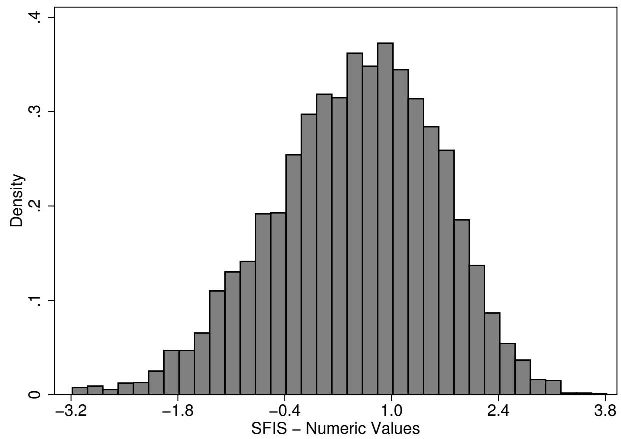


Figure 5. Distribution of SFIS-Numeric Scale

collaborative learning and student-faculty interaction benchmarks (Carini et al., 2006; Nelson Laird et al., 2008). Since the twelve questions used for the two scales are asked using both vague quantifier scales and z-scored numeric translations, four total scales are calculated for use: two for vague quantifier scales and two for z-scored numeric translations. These calculations were done by summing the responses for the vague quantifier scale and separately summing the numeric responses. In order to ensure comparability, all analyses comparing these scales include only cases where respondents gave answers to both measures, vague and numeric, for all questions in a scale. Means for each scale and the standard error (accounting for clustering) are presented in Table 1.

The distributions for these scales are presented in histograms in Figures 2-5. As can be seen, vague and numeric scales appear approximately normal. Both the ACLS and SFIS numeric scales have slightly more values at the lower end of the distributions, but overall have approximately similar distributions. Examining the data also suggests that there does not appear to be an issue with response styles in selection of the vague quantifiers. Only 0.6% of respondents selected the same vague quantifier for each of the behaviors. Conversely, 46.8% of respondents selected each of the response options at least once across all questions (i.e. used all options available) and another 45.4% selected 3 of the 4 response options at least once across all questions.

Different methods may be used to assess the comparative predictive validity of the measurement of these scales. The first is comparison of correlations between the outcome measures and differing measurement of the scales. Given the ranked ordering of the three outcome measures, grades, overall college satisfaction, and same college preference, Spearman's rho is the appropriate correlation coefficient to employ. The correlations of the two vague quantifier and two numeric scales with these two measures are presented in Table 2. Correlations are transformed to Fisher z-scores in order to examine significant differences. Due to the clustering effects, this may not be the most appropriate manner for testing of sig-

Table 1
Active-Collaborative Learning (ACLS) and Student-Faculty Interaction (SFIS) Scale Means

	ACLS		SFIS	
	Vague	Numeric	Vague	Numeric
Scale Mean	9.86	1.17	6.53	0.47
Std. Err.	0.08	0.02	0.07	0.02
n	9152		9474	

Table 2
Correlation of ACLS and SFIS with Grades and Satisfaction Measures

	Grades	College Satisfaction	Same College
Active-Collaborative Learning Scale			
ACLS-Vague	0.18*	0.24*	0.16*
ACLS-Numeric	0.10	0.17	0.13
n	9143	9150	9152
Student-Faculty Interaction Scale			
SFIS-Vague	0.15*	0.27*	0.19*
SFIS-Numeric	0.10	0.19	0.15
n	9463	9472	9470

* Significantly larger at $p < 0.05$ compared to other version of scale for same outcome measure

nificant differences; however, no other method is currently known, and this test gives a possible indication of significant differences. Further, in addition to being a test in comparative predictive validity used elsewhere (Al Baghal, 2011; Diefenbach et al., 1993), correlations have been used to test the predictive validity of NSSE benchmarks (e.g Carini et al., 2006). In any case, the point estimates for the correlation coefficients are not affected by the clustering, and are interpretable in terms of size and direction.

The results presented in these tables point in a single direction, being that vague quantifier responses have higher levels of predictive validity than do numeric open-ended responses. For both the ACLS and SFIS, for all three theoretically related variables, the point estimates for the correlations are larger for vague quantifier responses than for numeric responses. Further, all of these differences are statistically significant at the $p < 0.05$ level using the Fisher z-score transformation. Similarly small correlations are found in Carini et al. (2006), so these sizes are expected. Although the correlations as a whole are not large, what is more important is the difference in size of correlations between the two methods to measure the scales, particularly when this difference is significant. In particular, the effect size is always larger for vague quantifier scales than its numeric counterparts.

Another way to test the predictive validity of the vague quantifier and numeric responses scales is to use regressions predicting the theoretically related variables, while including important control variables. Since the purpose is to compare models in regards to which scale best predicts the theoretically related variables in order to further assess the predictive validity, not all possible control variables predicting these outcomes are required. The control variables selected, based on those used in Nelson Laird et al. (2008), are class standing (i.e. freshman, senior, or other), gender, full-time attendance status, and age (categorized as 19 and younger, 20-23, 24-29 and 30 and older).

The purpose of these regression models is to identify which scales increase model fit and predictive capability. Since the outcome variables (grades, overall college satisfaction, and same college preference) are ordinal-level variables, ordered logistic regressions are used taking into account the clustering effect of respondents within universities. Therefore, separate models for each of the scales predicting each of the three outcome variables are compared using the criterion of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) (Agresti, 2002). Since the AIC and BIC are related to the sample size, all models are restricted to include only cases where data is available for all measures within a model to ensure comparability. Thus, the number of cases is exactly the same for ACLS or SFIS models, but not necessarily the same across scales. A base model is also estimated for each of the three outcome variables. This model included only the demographic variables to show improvement in the models based on the inclusion of the various scales. The AIC and BIC both are indicators of relative quality of statistical models, taking into account the goodness-of-fit and level of complexity (i.e. number of parameters) in the model. Lower AIC and BIC indicate preferable models in terms of fit given complexity. AIC and BIC do not reflect tests of null hypotheses per se, and the only criteria for choice is the lowest value estimated. The AIC and BIC for each of the outcome variables and each of the scales are presented in Table 3 (decimals rounded to increase clarity of presentation).

The results mirror those of the correlation analyses, with the models employing the vague quantifier versions of the scales performing better than either the base model or the model using the numeric versions of the scales, as indicated by the lower AIC and BIC scores. These findings hold true in all cases for both the active-collaborative learning scale and for the student-faculty learning scale when examining either the AIC or BIC. The models are restricted to the same sample and include all of the same control variables, only differing in which version of the scale (numeric or vague quantifier version) are used. Unlike the correlation analyses, the estimated variances take into account the clustering identified in the data. Therefore, the differences in the AIC and BIC can be solely attributed to the differential predictive validity for each

Table 3
Logistic Regression Indicators of ACLS and SFIS on Grades and Satisfaction Measures

	Grades		College Satisf.		Same College	
	AIC	BIC	AIC	BIC	AIC	BIC
Active-Collaborative Learning Scale						
Base	32666	32766	18402	18474	20689	20761
ACLS-Vague	32432	32539	17837	17915	20404	20483
ACLS-Numeric	32555	32662	18131	18209	20529	20607
n	9119		9124		9123	
Student-Faculty Interaction Scale						
Base	33711	33811	19056	19128	21381	21452
SFIS-Vague	33548	33656	18321	18400	20989	21067
SFIS-Numeric	33594	33702	18659	18738	21115	21194
n	9439		9446		9445	

Abbreviations: AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion

Table 4
Standardized Coefficients and Odds Ratios of ACLS, SFIS on Grades and Satisfaction Measures

	Grades			College Satisfaction			Same College		
	Coef.	S.E.	OR	Coef.	S.E.	OR	Coef.	S.E.	OR
Active-Collaborative Learning Scales									
ACLS-Vague	0.166*	0.007	1.091	0.286*	0.006	1.163	0.196*	0.007	1.109
ACLS-Numeric	0.113*	0.025	1.266	0.188*	0.026	1.484	0.140*	0.027	1.340
n		9119			9124			9123	
Student-Faculty Interaction Scale									
SFIS-Vague	0.133*	0.006	1.080	0.320*	0.008	1.224	0.226*	0.008	1.158
SFIS-Numeric	0.112*	0.016	1.190	0.225*	0.021	1.428	0.179*	0.019	1.301
n		9439			9446			9445	

Abbreviations: Coef. = Coefficient; S.E. = Standard Error; OR = Odds Ratio
* $p < 0.05$

version of the scales. As such, it appears that vague quantifiers have higher levels of predictive validity than scales measured using numeric open-ended responses.

Finally, the estimated standardized coefficients (to allow comparison) and odds ratios for the different scales in predicting the three dependent variables controlling for demographics are presented in Table 4. The standard errors reported reflect the clustered sample design using the Taylor series approximation. All of the coefficients are significant, and all coefficients are in the expected direction, with increases in the scales being associated with increases in the dependent variables. That is, increases in both the active-collaborative learning scale and student-faculty interaction scale, regardless of scale type (vague or numeric), lead to increases in the predicted grade and satisfaction with college experiences.

In addition, all of the standardized coefficients are larger for the vague quantifier versions of the scales than for the numeric versions in the three comparisons. Taken together, this difference in effect size suggests that changes along the vague quantifier scales have more influence in the predicted outcomes of grades and satisfaction with college experience than do numeric versions of the same scales. This finding further supports the position that vague quantifiers display higher levels of predictive validity than numeric responses. While not substantively large differences exist across all these effect sizes or the correlations, some larger differences are evident, particularly for college satisfaction. Further, all the differences are significant, suggesting that there is an improvement in measurement using vague quantifiers, which may be contrary to the arguments of many researchers, and is suggestive for choices and possible improvements in survey design.

4 Discussion and Conclusions

This study compared predictive validity of responses to vague quantifier and numeric open-ended questions, which is important in determining which of the two question types has better measurement properties (Chang & Krosnick, 2003). Many survey researchers have expressed reservations about the use of vague quantifiers as a measurement tool. These reservations have extended into research in higher education. Questions about the validity of the student surveys, and in particular the NSSE, have focused on a number of reasons that the data may be problematic, including the measurement properties of vague quantifiers. This research examines the efficacy of vague quantifiers to measure behavioral frequencies compared to other possible measures, which is important to a range of surveys, including educational surveys such as the NSSE. No study to date has examined the comparative predictive validity of these two types of measures for frequency estimation. The closest to examining response differences is Lu et al. (2008); however, these researchers use an uncommon variation of frequency estimation (percentage scales) rather than the frequently used open-ended format. These numeric scales also may be problematic for other reasons (e. g. Schwarz et al., 1985).

Unlike the Lu et al. (2008) study, this research finds significant differences between response formats. Specifically, vague quantifier responses display higher levels of predictive validity than those using numeric open-ended responses. This higher level of predictive validity holds regardless of which scale is inspected (i.e. active-collaborative learning scale or student-faculty interaction scale). It also does not matter which outcome variable (grades, satisfaction with college experience, or same college preference) these scales are correlated with or predicting. Interestingly, these results hold up if controlling for possible measures of numeracy. Additional analysis categorizing respondents based on their Scholastic Aptitude Test (SAT) Math scores available in the NSSE found that in all categorizations, vague quantifiers performed better than numeric responses (not shown).[4] The lack of differences suggests the possibility that the effectiveness of vague quantifiers relative to numeric responses is not affected by numerical ability.

Taken as whole, this study increases the understanding of numeric quantity measurement in surveys, suggesting that vague quantifiers may have better measurement properties than numeric responses, at least in some ways. These findings add to several other studies also showing the possible value of vague quantifiers in terms of predictive validity (Al Baghal, 2011; Windschitl & Wells, 1996) and accuracy (Al Baghal, 2014). In light of such findings, the current study adds further evidence that vague quantifiers may be useful measures in student surveys and surveys more generally, especially if the goal is to examine potentially important relationships between variables. In the case of the use of the NSSE among higher learning institutions, it is indeed the relationship between these scales and performance outcomes that are of particular importance.

Vague quantifiers may have performed better in terms of predictive validity for a number of reasons. First, as noted, people generally may not think in numerical terms. Second, the relative component of vague quantifiers may capture aspects of the behavior that allow respondents to understand their behavior in relation to others (Schaeffer, 1991). Students may know that they performed some action a given number of times, but the comparison to other students provides context to understand the relative importance of their actions. Similarly, this context and other contextual factors, such as perceived importance, may influence respondents' selection of the vague quantity beyond just the frequency (Moxey & Sanford, 2000). The relativity and importance vague quantifiers communicate about behaviors may increase its discrimination between students and the influence of measures on other important outcomes.

Another finding of this study, although not necessarily related to the use of vague quantifier or numeric responses, is the effects of clustering. Overall, the mean design effect approached 4 (mean = 3.888), indicating a near four-fold increase in variance due to clustering compared to the usual simple random sampling assumptions. Although clustering effects are frequently identified in educational research, with students more similar within institutions, it is worth noting as although clustering effects can occur in web surveys, clustering effects are most frequently thought of occurring in face-to-face surveys, in part due to neighborhood and interviewer effects (Lohr, 2010).

Although this study took appropriate steps to estimate variances correctly, there are some possible limitations to this research. The first is that the vague quantifier versions of the questions were always asked first; the context these questions were asked in were therefore not randomly assigned and controlled. Another is that given the differences in scale construction, a relatively infrequent occurring event in the numeric scale may have relatively less impact than a selection of a smaller vague quantifier (e. g. sometimes) would have on the vague quantifier scale. The numeric data was standardized, which is felt alleviates most of the problem; a similar transformation was used and found that the relationship between vague quantifier response and numeric representation approximated linearity (Nelson Laird et al., 2008). Still, it is possible differences in relative importance influenced results, and should be examined in further studies using numeric translations.

Further, all respondents are college students, meaning that

[4] SAT Math scores are not available for all institutions, since some universities decline to provide or do not require the SAT from their students. Hence this analysis was only possible on 4761 respondents.

for the most part, the respondents are of a certain age range. The use of this population limits the potential generalizability somewhat to survey research more generally or student surveys in lower age-ranges, for example. However, it may be expected that college students are more numerate, and hence, the effects found here may be more strongly identified in a general population who are less able to use numeric information. Additional attention to the effect of numeracy is warranted, including use of measures better than SAT scores.

Acknowledgements

The author would like to thank the Indiana University Center for Postsecondary Research for their permission to use their experimental data set. In addition, the author wishes to thank the reviewers for suggestions that greatly improved the manuscript.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: Wiley.

Al Baghal, T. (2011). The measurement of risk perceptions: the case of smoking. *Journal of Risk Research*, *14*(3), 351–364.

Al Baghal, T. (2014). Numeric estimation and response options: an examination of the accuracy of numeric and vague quantifier responses. *Journal of Methods and Measurement in the Social Sciences*, in press.

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

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

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

Burton, S. & Blair, E. (1991). Task conditions, response formulation processes and response accuracy for behavioral frequency questions in surveys. *Public Opinion Quarterly*, *55*, 50–79.

Carini, R. M., Kuh, G. D., & Klein, S. P. (2006). Student engagement and student learning: testing the linkages. *Research in Higher Education*, *47*, 1–32.

Chang, L. & Krosnick, J. (2003). Measuring the frequency of regular behaviors: comparing the "typical week" to the "past week". *Sociological Methodology*, *33*, 55–80.

Cole, J. S. & Korkmaz, A. (2013). Estimating college student behaviour frequencies. *Journal of Applied Research in Higher Education*, *5*, 58–71.

Diefenbach, M. A., Weinstein, N. D., & O'Reilly, J. (1993). Scales for assessing perceptions of health hazard susceptibility. *Health Education Research*, *82*, 181–192.

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

Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: overcoming low numeracy. *Health Psychology*, *28*, 210–216.

Gellin, A. (2003). The effect of undergraduate student involvement on critical thinking: a meta-analysis of the literature, 1991–2000. *Journal of College Student Development*, *44*, 746–762.

Kish, L. (1965). *Survey sampling*. New York: Wiley.

Kuh, G. D. (2003). What we're learning about student engagement from NSSE: benchmarks for effective educational practices. *Change*, *35*, 24–32.

LaNasa, S. M., Olson, E., & Alleman, N. (2007). The impact of on-campus student growth on first-year student engagement and success. *Research in Higher Education*, *48*(8), 941–966.

Lohr, S. L. (2010). *Sampling: design and analysis* (2nd ed.). Pacific Grove: Duxbury Press.

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.

Moxey, L. M. & Sanford, A. J. (2000). Communicating quantities: a review of psycholinguistic evidence of how expressions determine perspectives. *Applied Cognitive Psychology*, *14*, 237–255.

National Survey of Student Engagement. (2006). NSSE 2006 overview. Retrieved from http://nsse.iub.edu/pdf/2006_Institutional_Report/NSSE%202006%20Overview.pdf

National Survey of Student Engagement. (2013). About NSSE. Retrieved from http://nsse.iub.edu/html/about.cfm

Nelson Laird, T. F., Korkmaz, A., & Chen, D. (2008). *How often is "often" revisited: the meaning and linearity of vague quantifiers used in the National Survey of Student Engagement.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego.

Pace, R. C. & Friedlander, J. (1982). The meaning of response categories: how often is "occasionally," "often," and "very often"? *Research in Higher Education*, *17*(3), 267–281.

Porter, S. R. (2011). Do college student surveys have any validity? *The Review of Higher Education*, *35*, 45–76.

Revilla, M. & Saris, W. (2012). A comparison of the quality of questions in a face-to-face and a web survey. *International Journal of Public Opinion Research*, *25*, 242–253.

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.

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

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

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: Cambridge University Press.

Weinstein, N. D. & Diefenbach, M. A. (1997). Percentage and verbal category measures of risk likelihood. *Health Education Research*, *121*, 139–141.

Windschitl, P. D. & Wells, G. L. (1996). Measuring psychological uncertainty: verbal versus numeric methods. *Journal of Experimental Psychology: Applied*, *2*, 343–336.

Appendix A
Active and Collaborative Learning Scale

In your experience at your institution during the current school year, about how often have you . . .

☐ Asked questions in class or contributed to class discussions
☐ Made a class presentation
☐ Worked with other students on projects during class
☐ Worked with classmates outside of class to prepare class assignments
☐ Tutored or taught other students (paid or voluntary)
☐ Participated in a community-based project (e.g., service learning) as part of a regular course
☐ Discussed ideas from your readings or classes with others outside of class (students, family members, co-workers, etc.)

Appendix B
Student-Faculty Interaction Scale

In your experience at your institution during the current school year, about how often have you . . .

☐ Discussed grades or assignments with an instructor
☐ Talked about career plans with a faculty member or advisor
☐ Received prompt written or oral feedback from faculty on your academic performance
☐ Worked with faculty members on activities other than coursework (committees, orientation, student life activities, etc.)
☐ Discussed ideas from your readings or classes with faculty members outside of class

Appendix C
Theoretically Related Variables of Interest

What have most of your grades been up to now at this institution?

☐ C- or lower
☐ C
☐ C+
☐ B-
☐ B
☐ B+
☐ A-
☐ A

How would you evaluate your entire educational experience at this institution?

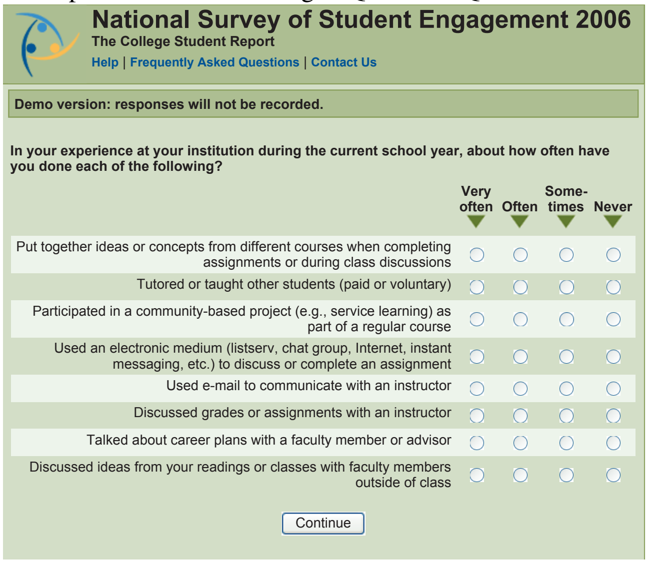
☐ Poor
☐ Fair
☐ Good
☐ Excellent

If you could start over again, would you go to the *same institution* you are now attending?

☐ Definitely no

☐ Probably no
☐ Probably yes
☐ Definitely yes

Appendix D
Example Screen of How Vague Quantifier Questions Asked


National Survey of Student Engagement 2006
The College Student Report
Help | Frequently Asked Questions | Contact Us

Demo version: responses will not be recorded.

In your experience at your institution during the current school year, about how often have you done each of the following?

	Very often	Often	Some-times	Never
Put together ideas or concepts from different courses when completing assignments or during class discussions	○	○	○	○
Tutored or taught other students (paid or voluntary)	○	○	○	○
Participated in a community-based project (e.g., service learning) as part of a regular course	○	○	○	○
Used an electronic medium (listserv, chat group, Internet, instant messaging, etc.) to discuss or complete an assignment	○	○	○	○
Used e-mail to communicate with an instructor	○	○	○	○
Discussed grades or assignments with an instructor	○	○	○	○
Talked about career plans with a faculty member or advisor	○	○	○	○
Discussed ideas from your readings or classes with faculty members outside of class	○	○	○	○

Continue