Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

# Question-order effects in social network name generators[☆]

James E. Pustejovsky*, James P. Spillane

Northwestern University, School of Education and Social Policy, 2120 Campus Drive, Evanston, IL 60201, USA

ARTICLE INFO

*Keywords:*
Social network name generators
Reliability
Survey design

ABSTRACT

Social network surveys are an important tool for empirical research in a variety of fields, including the study of social capital and the evaluation of educational and social policy. A growing body of methodological research sheds light on the validity and reliability of social network survey data regarding a single relation, but much less attention has been paid to the measurement of multiplex networks and the validity of comparisons among criterion relations. In this paper, we identify ways that surveys designed to collect multiplex social network data might be vulnerable to question-order effects. We then test several hypotheses using a split-ballot experiment embedded in an online multiple name generator survey of teachers' advice networks, collected for a study of complete networks. We conclude by discussing implications for the design of multiple name generator social network surveys.

Social network surveys are an important tool for empirical research in a variety of disciplines and applied fields, including the study of social capital and the evaluation of educational and social policy. A growing body of methodological research sheds light on the validity and reliability of measurements of a single relationship among a set of actors. However, much less attention has been paid to the measurement of multiplex networks and the validity of comparisons between criterion relations, despite the fact that many research questions require attention to several types of relationships among a given set of actors.[1] In this paper, we identify ways that surveys designed to collect multiplex social network data might be vulnerable to question-order effects, then test several hypotheses using a split-ballot experiment embedded in an online multiple name generator survey.

Among the many design choices involved in constructing a multiplex network survey, the researcher is faced with the question of how the measurement of several relational criteria should be arranged. For example, which relation should be measured first? The possibility that the order in which questions are posed could create bias, or what we term here a question-order effect, is an immediate concern for researchers relying on survey designs (Burt, 1997; Ruan, 1998; Straits, 2000).

In order to measure multiplex networks, survey methods for measuring single networks, which include roster-based methods and recall-based methods, are typically extended to cover multiple criterion relationships. In organizational network studies where the set of all relevant actors can be determined in advance, roster-based recognition methods can be applied. Through binary or rating-scale questions, roster-based surveys ask a respondent to specify, classify, or characterize their relationship with each member of a pre-determined group (Marsden, 2005). The researcher may specify relationships in terms of a hypothetical criterion, factual criterion, or a semantic differential (De Lange et al., 2004). However, roster methods may present a considerable reporting burden if participants are asked to report on each member of a large organization. In some cases, the reporting burden may be lessened by limiting the survey to a sub-set of names, chosen based on the organizational structure (Reagans and McEvily, 2003).

To measure multiplex networks using rosters, questions would be posed about each of several criterion relationships. Questions are posed either question-wise, where the respondent answers one criterion relationship question about the entire set of possible alters before repeating the process with the next criterion relationship,

[☆] An earlier version of this paper was presented at the 28th annual International Sunbelt Social Network Conference, January 22–28, 2008 in St. Pete Beach, Florida. Work on this project is supported by the Distributed Leadership Studies (http://www.distributedleadership.org), funded by research grants from the National Science Foundation (RETA Grant EHR-0412510), with support from Northwestern University's School of Education and Social Policy and Institute for Policy Research. We thank Michael Lach, Gina Grant, Bill Suttles, Ruth Heaton, and Jim Lewis for their assistance with data collection, and Ken Frank for many helpful comments on our analysis. We also gratefully acknowledge the helpful feedback we received from two anonymous reviewers, and the dedicated support of our research team, especially Enrique Orlina, Virginia Pitts, Cindy Sigal, Bijou Hunt, and Cassandra Chen. All opinions and conclusions expressed in this paper are our own and do not reflect the views of any funding agency.

* Corresponding author.
*E-mail address:* j-pustejovsky@northwestern.edu (J.E. Pustejovsky).

[1] For example, in studies of personal social capital, multi-dimensional network data have been used to examine differences between emotional support networks and social support networks (Bernard et al., 1990), to identify factors affecting reciprocal exchange of support (Plickert et al., 2007), and to validate widely used name generator questions (Ruan, 1998; De Lange et al., 2004). In the realm of organizational network analysis, multi-dimensional networks have been used to study patterns in the relational structure of a law firm spread across multiple offices (Lazega and Pattison, 1999), to determine the dimensionality of advice seeking behavior (Cross et al., 2001), and to identify factors related to the career advancement of managers in large corporate firms (Burt, 1997; Podolny and Baron, 1997).

0378-8733/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.socnet.2009.06.001

or alter-wise, where a respondent characterizes her relationship with one possible alter in terms of all the criterion relationships of interest before moving on to the next possible alter (Vehovar et al., 2008; Kogovšek et al., 2002).

Recall-based methods use name generator questions, which ask the respondent to name a set of people that fit a given criterion relationship. The criterion relationship can be formulated in various ways: by specifying a social role, a minimum frequency of contact, closeness, or a specific type of social exchange (Marin and Hampton, 2007; van der Poel, 1993). Recall-based methods are often necessary because all of the relevant possible members of a network cannot be identified in advance, thereby preventing the use of roster methods. A name generator may be followed by a set of name interpreter questions that ask the respondent to provide additional information about some or all of the individuals they have named.

To measure multiplex networks using recall methods, a sequence of two or more name generators would be posed; each name generator would ask the respondent to list people that fit a specific criterion relationship. Name interpreter questions might also be posed about the contacts from each name generator, or about the total set of unique contacts mentioned in any of the name generators.

In this paper we focus on such recall-based approaches for measuring multiplex social networks, examining the effects of question-order on responses to multiple name generator surveys. After briefly reviewing previous research on measuring social networks, we outline several ways in which surveys that collect multiplex social network data might suffer from measurement bias. We test our theories using a randomized field experiment embedded in two studies of complete advice networks among elementary and middle school teachers. We conclude by drawing implications for the design of survey instruments to measure multiplex social networks, arguing that much of our work is applicable not just to complete network studies, but to ego-centric designs as well.

## 1. Question-order effects in social network name generators

A sizable literature addresses the accuracy and reliability of name generator questions (Marsden, 2005 provides an excellent overview), but most of this work focuses only on networks defined on a single relationship, such as acquaintanceship. In the context of single-network studies, researchers have criticized recall-based network data as inaccurate (Bernard and Killworth, 1977) and demonstrated that respondents often forget to report alters (for a review, see Brewer, 2000). Others have found that accuracy is mediated by network size (Bell et al., 2007), the specificity and salience of the name generator (*ibid.*), the closeness, frequency of contact, and recency of contact with alters (Hammer, 1984), and the degree to which respondents have elaborated a mental framework for remembering interactions or relationships that they are asked to recall (Freeman et al., 1987).

Multiplex network data present an additional set of validity concerns, particularly for studies comparing one criterion relation to another, yet there is a scarcity of work that studies multiple name generator survey designs. Ferligoj and Hlebec (1999) demonstrated question-order effects on reliability, concluding that network data from later name generators is somewhat more reliable than data from initial name generators. Examining the drawbacks of a sequential, multiple name generator survey design, Straits (2000) suggested three possible mechanisms that would produce question-order effects: priming, fatigue, or a reluctance to repeat alters for fear of being redundant. Below we elaborate on these mechanisms and several others, drawing from research on cognitive aspects of survey methodology.

Studies of question-order effects in behavioral and attitudinal surveys suggest specific mechanisms that may also be applicable to multiple name generator social network surveys. Most research on the cognitive aspects of survey methodology focuses on attitude questions, which ask respondents to select an opinion from a list of options or evaluate their level of agreement with a statement, or on behavioral frequency questions, which ask respondents to report on how often they have engaged in certain behaviors (Sudman et al., 1996, Tourangeau; for an overview of research in cognitive aspects of survey methodology, see Sirken et al., 1999). Though a social network name generator question presents a very different set of concerns than the Likert-scale item design that is typical in attitudinal surveys, we believe that many of the same cognitive principles may apply.

Question-order effects might appear as a result of at least five inter-related mechanisms: fatigue, satisficing, conversational norms of non-redundancy, cognitive priming, or question-scope redefinition (Tourangeau and Rasinski, 1988; Straits, 2000). We take up each of these mechanisms in turn, examining conditions under which they might cause question-order effects in multiple name generator surveys, and reasoning about the resulting bias.

The following examination of question-order effects is limited to those that appear as a result of using sequential name generators in a survey. For explanatory purposes, we imagine a survey with two name generators. Question-order effects are present to the extent that the network as measured by the second name generator differs substantially from the network that would have been produced by the second name generator, if the first name generator were not asked. (By extension, in a survey containing three or more name generators, question-order effects are present to the extent that the network as measured by a given name generator differs from the network that would have been produced, were any of the preceding name generators not posed.)

### 1.1. Fatigue

Respondent fatigue is perhaps the simplest mechanism that could create question-order biases. Fatigue effects create bias if, in response to the second name generator, a respondent names fewer alters than she otherwise would have, had the first name generator not been posed. In the extreme, fatigue might lead to non-response to later name generators and name interpreters. Fatigue effects may be particularly pronounced in surveys where the overall length depends on the number of items named in response to a question (Tourangeau and Rasinski, 1988). As discussed below, the name interpreter questions in our survey create such a situation.

Fatigue effects would lead to a diminished average out-degree, a pattern of bias that would be particularly troublesome for multiplex network studies seeking to compare the relative size or density of two networks. Beyond this basic measure, fatigue effects could create bias that is connected with patterns in the reported order of alters, if alters that would typically be reported further down the list are censored. For example, if alters that are encountered less frequently tend to be reported later down the list in response to a given name generator, the respondent's average frequency of interaction would be biased upward.

### 1.2. Satisficing

Satisficing effects occur when a respondent, perhaps due to fatigue, boredom, or confusion, gives a response that she believe satisfies the request for information, but is not a complete, optimally considered response (Krosnick, 2000). The theory of satisficing as applied to attitude questions is used to explain primacy or recency effects, acquiescence bias, and status-quo or no-opinion bias (Krosnick et al., 1996). Satisficing behavior is thought to be regulated by task difficulty, respondent ability, and respondent motivation (*ibid.*).

In the context of name generator prompts, satisficing would play a role as a respondent decides how many alters to list in response to a name generator prompt. A satisficing respondent will take cues from the design of the survey to determine the number of names necessary for a sufficient response to a name generator. Such cues might include the number of lines provided after a name generator prompt, which a respondent takes as an indication of the researcher's expectation about the range of items that will be listed (Vehovar et al., 2008). In a multiple name generator survey, the first prompt in a survey is a novel question, but the second prompt follows the same pattern as the first. When confronted with the second name generator prompt, a satisficing respondent could turn to the precedent that she herself set by responding to the first name generator. For example, she may stop searching her memory after listing three names in the second generator, because the three names that she listed in the first generator seemed to be an adequate response.

Fatigue effects and satisficing effects are competing theories of how respondents answer survey questions when tired or under-motivated. Fatigue effects lead to downward bias in out-degree, regardless of the relative size or density of the networks being measured. In contrast, satisficing leads to downward bias in measured out-degree only to the extent that a respondent has a higher (actual) out-degree in the second network than in the first. If the reverse is true, satisficing produces no bias in the network, or might even lead a respondent to list more names than she otherwise would have, so as to match the precedent set in response to the first name generator.

### 1.3. Non-redundancy

Studies of attitudinal surveys have found that respondents sometimes interpret subsequent questions as requests only for new information rather than as independent questions, leading them to omit consideration of information that they have already offered (Schwarz, 1999). In the context of multiple name generator surveys, non-redundancy effects would appear if a respondent omits the names of certain alters in the second name generator because she has already listed those alters in the first name generator. The respondent might interpret the second name generator prompt as beginning with the qualification, "Aside from the people you have already named. . ."

Non-redundancy effects are difficult to observe, because in advance of measurement it is difficult to know how much overlap is present in a given pair of networks or relationships.[2] Non-redundancy effects are observed if a respondent's relationship to a given alter fits the criteria specified for both name generators (i.e., the respondent has multiplex ties to the alter), but the respondent names the alter only in the first name generator. The bias created by non-redundancy effects therefore depends on the actual prevalence of multiplex ties. The overall effect of non-redundancy is to reduce the average out-degree (and density) in the second network and to reduce the in-degree of actors that are a part of both networks (thus biasing downward the observed level of multiplexity), though all of these effects will only be present to the extent that that the two networks overlap.

### 1.4. Cognitive priming

Cognitive priming in an attitude or behavioral question affects the retrieval of information from memory that is relevant to answering the question (Tourangeau and Rasinski, 1988). A previous question may have sub-consciously activated a set of relevant memories. These memories would not otherwise have been drawn upon in forming a response to the current question, but now their activation could cause a change in the response.[3] In the context of a multiple name generator survey, the process of retrieving names from memory for the first generator may start a sub-conscious activation process that brings certain names to the forefront for subsequent name generator questions. If not for the priming effect of the earlier name generator, a respondent might not have listed certain alters in the current name generator.

Priming does not necessarily bias the number of names listed in response to name generators; rather, it produces bias by changing the set of names that a respondent considers when determining which of her relationships fit the criterion specified in the generator. Priming would have a greater effect on the results of a second name generator to the extent that the second network is composed of a different set of actors than the first. If the central actors in both networks are largely distinct, then priming the actors in the first network would result in additions to (or perhaps replacement of) the set of alters named in the second network. If the actors in both networks are largely the same, then priming the actors in the first network might have little effect on, or might even increase the accuracy of, the alters named in the second network. Priming effects therefore create bias that is directionally opposed to non-redundancy effects, by increasing the similarity between networks measured with subsequent name generators.

### 1.5. Question-scope redefinition

Question context effects result from the manner in which survey respondents rely on the wording of specific questions, the sequencing of questions (adjacent questions, in particular), and other facets of the instrument to infer the pragmatic meaning of a question (Schwarz, 1999). Social network name generators are no exception; a respondent must make some assumptions about the sort of names that a name generator question is intended to produce, and will look for contextual clues in order to understand the relationship being described (Bailey and Marsden, 1999).[4] If a respondent relies on contextual clues from the first name generator to understand the pragmatic meaning of the second generator, the alters that she names may be different from those she would have named in the absence of the first generator. For example, asking "Please list the names of five friends" as an initial name generator may produce a wide variety of responses, because the "friend" criterion is fairly ambiguous. If instead the "friends" name generator is preceded by questions about childhood experiences, the scope of the name generator may be implicitly re-defined to focus exclusively on childhood friends.

[2] In fact, measuring network multiplexity is sometimes a substantive question for study; see for instance Ruan (1998) and Lazega and Pattison (1999).

[3] For example in the context of behavioral frequency questions, one experiment found that first answering a set of questions about one's general opinions of crime and victimization led to increased reporting of victimization incidents in the past year (Cowan et al., 1978).

[4] McPherson et al. (2006) also consider question-scope redefinition as a possible confounding effect in their comparison of ego-centric network size from the 1985 and 2004 General Social Surveys.

Question-scope redefinition would produce empirical effects that are very similar to those produced by priming. To the extent that a respondent assumes that the meaning of the first name generator is similar to the meaning of the second, the results from the second name generator should more closely resemble the results from the first name generator. Conceptually, question-scope re-definition could be distinguished from priming based on the respondent's cognitive process. Priming occurs at the level of subconscious memory processes, whereas question-scope redefinition has to do with respondent's understanding and interpretation of the question, something that they should be able to express.

The five possible sources of question-order effects that have been identified fall into three areas, which structure our empirical analysis below. Fatigue and satisficing effects act most directly on out-degree, the number of alters that a respondent lists. They offer competing hypotheses regarding the direction of bias. Non-redundancy and priming effects are directly related to the amount of overlap between networks defined on different criterion relationships. Our experimental design lets us say very little about the extent of these effects. Finally, question-scope redefinition can potentially produce biases in both name generators and name interpreters. Below we examine empirical evidence of question-order effects, using a split-ballot experiment embedded in an online social network survey.

## 2. Survey design and research methods

The question-order experiment was embedded in a larger survey designed to study social capital in elementary and middle schools by measuring advice relationships among school staff.[5] Prior research led us to recognize that the school subject is an important consideration in the structure of social relations among school staff (Hayton and Spillane, 2007; Burch and Spillane, 2005; Drake et al., 2001); consequently, we decided to collect multiplex network data, differentiating advice networks by school subjects. We use a split-ballot experimental design to test whether the order of name generator prompts in the survey affects the validity of inferences made based on the resultant multiplex network data. In this section, we explain the relevant aspects of the instrument design, describe two studies that made use of the instrument, and outline our approach to data analysis.

### 2.1. Instrument design

One portion of the survey consists of a sequence of name generators and interpreters. Each name generator begins with the same wording: "In the past year, to whom have you gone for advice or information about teaching [SUBJECT PROMPT]?" Each name generator is followed by a series of name interpreter questions. For every alter that a respondent names, data is collected on the role or job description of the alter, the content of the advice interactions, the frequency of interactions between respondent and alter, and the respondent's rating of the influence of the alter's advice on her work.[6] We created an experimental mechanism in the survey by randomizing the order of the name generator prompts. Respondents have a 50% chance of receiving the math name generator and interpreters first, followed by the reading name generator and interpreters, and a 50% chance of receiving the same prompts, but in the opposite order.

**Table 1**
Number of respondents by treatment.

| Sample | Treatment | | Total |
|---|---|---|---|
| | Reading First | Math First | |
| Elementary schools | 105 | 107 | 212 |
| Middle schools | 48 | 37 | 85 |
| Combined sample | 153 | 144 | 297 |

### 2.2. Data collection

In the analysis that follows, we use results from two samples that were collected using the same instrument. We limit our analysis to the sub-set of respondents who report teaching both mathematics and reading/writing/Language Arts.[7] The first sample consists of 15 public elementary schools and 4 Catholic elementary schools (most serving kindergarten through 8th grade) in a large U.S. city. School faculties vary in size from 14 to 69. All teachers, administrators, and school-level specialists were asked to complete the web-based survey during a 6-week period in the Spring of 2007. In this sample, we received a full or partial response from 414 out of 544 staff (76%); of these, 212 were from respondents who teach both math and reading.[8] This portion of the sample is composed mostly of contained-classroom, primary grade teachers. Table 1 presents the number of respondents by sample and treatment group.

The second sample consists of 10 public middle schools in a mid-sized city in a different state, all serving grades 6 through 8. All teachers, administrators, and school-level specialists were asked to complete the web survey. School faculties range in size from 49 to 69 certified staff. We received a full or partial response from 548 out of 634 staff (87%); of these, 85 were from respondents who teach both reading and math.[9] This portion of the sample is composed entirely of sixth grade teachers in self-contained classrooms.

### 2.3. Data analysis

The survey design permits us to detect question-order effects by comparing the two randomly assigned sub-sets of each sample. Approximately half of our respondents answered the math name generator and interpreter questions before the reading name generator; below we refer to this group as the Math First treatment. The remaining respondents answered the reading name generator and interpreter questions before the math name generator; we refer to them as the Reading First treatment. For each subject area, we compare the data from the treatment where the name generator was posed first to the data from the treatment where the name generator was posed second. We assume that assigning the two treatments groups creates a random partition of the out-degree distribution, so that any differences between these distributions are attributable to question-order effects.

Throughout, we use non-parametric tests of significance. Network degree distributions are typically very skewed; often one observes that a few individuals have many ties to other, while most

[5] For a broad overview of the survey, see Pustejovsky et al. (2009). The full survey can also be accessed at the following website: http://www.sesp.northwestern.edu/Survey/SchoolNetworkSurvey.html.

[6] Note that the number of name interpreter items that a respondent is asked to answer depends on the number of names she lists in the name generator. The total length of the survey therefore depends in part on the length of a respondent's list of alters.

[7] In the remainder of this paper, we use the term reading to abbreviate Reading/Writing/Language Arts.

[8] School-level response rates range from 41% to 95%. The two treatment groups do not differ in mean age, mean years of teaching experience, percent female, or percent minority.

[9] The two treatment groups do not differ in mean years teaching experience or percent minority. The groups do however differ somewhat in mean age (40.6 year versus 44.0 years) and percent female (66% versus 84%), though these difference are not statistically significant at the 5% level.

**Table 2a**
Mean reading out-degree.

| Sample | Treatment group | | Difference (std. error) | Standardized *U* statistic[a] |
|---|---|---|---|---|
| | Reading First | Math First | | |
| Elementary schools | 2.65 | 1.32 | 1.33 (0.29) | 4.04[b] |
| Middle schools | 4.42 | 3.03 | 1.39 (0.49) | 2.92[b] |
| Combined sample[c] | | | 1.35 (0.25) | 3.75[b] |

[a] The Mann–Whitney *U* statistic tests the null hypothesis that the observed sample from each treatment group was drawn from a common distribution. Under the null hypothesis, the standardized Mann–Whitney *U* statistic is approximately normally(0,1) distributed.

[b] The null hypothesis is rejected at the 5% level.

[c] The estimates for the combined sample are calculated by taking the weighted average of the estimate from each sample, with weights inversely proportional to the variance of the estimated difference.

**Table 2b**
Mean math out-degree.

| Sample | Treatment group | | Difference (std. error) | Standardized *U* statistic[a] |
|---|---|---|---|---|
| | Reading First | Math First | | |
| Elementary schools | 1.22 | 1.13 | 0.09 (0.21) | 0.43 |
| Middle schools | 3.35 | 2.86 | 0.49 (0.49) | 0.90 |
| Combined sample[b] | | | 0.15 (0.20) | 0.51 |

[a] The Mann–Whitney *U* statistic tests the null hypothesis that the observed sample from each treatment group was drawn from a common distribution. Under the null hypothesis, the standardized Mann–Whitney *U* statistic is approximately normally(0,1) distributed.

[b] The estimates for the combined sample are calculated by taking the weighted average of the estimate from each sample, with weights inversely proportional to the variance of the estimated difference.

**Table 2c**
Mean (standard error) total out-degree.

| Sample | Treatment group | | Difference (std. error) | Standardized *U* statistic[a] |
|---|---|---|---|---|
| | Reading First | Math First | | |
| Elementary schools | 3.87 | 2.45 | 1.42 (0.46) | 3.00[2] |
| Middle schools | 7.77 | 5.89 | 1.88 (0.86) | 2.15[b] |
| Combined sample[c] | | | 1.52 (0.40) | 2.81[b] |

[a] The Mann–Whitney *U* statistic tests the null hypothesis that the observed sample from each treatment group was drawn from a common distribution. Under the null hypothesis, the standardized Mann–Whitney *U* statistic is approximately normally(0,1) distributed.

[b] The null hypothesis is rejected at the 5% level.

[c] The estimates for the combined sample are calculated by taking the weighted average of the estimate from each sample, with weights inversely proportional to the variance of the estimated difference.

other individuals have very few ties (Wong et al., 2006). Normality assumptions are likely to be invalid, making the use of *t*-tests inappropriate; Mann–Whitney tests provide an alternative that makes no distributional assumption.

## 3. Findings

The two samples reveal a consistent pattern of question-order effects. We first examine effects on out-degree, discussing the evidence for fatigue effects versus satisficing effects. We then turn to non-redundancy and primacy effects, and conclude by examining evidence of question-scope redefinition.

For both samples, the reading name generator reveals significant differences between treatment groups in the distribution of out-degree. Respondents in the Math First treatment, who received the reading name generator after first answering the math name generator, list an average of 1.33 fewer names than respondents in the Reading First treatment, a decrease of 50% in the number of names generated (see Table 2a). In contrast, results from the math name generator do not display a significant difference between treatment groups. In the elementary school sample, respondents in the Math First treatment list an average of 0.09 fewer names for mathematics than respondents in the Reading First treatment, a difference which is neither large in magnitude nor statistically significant (see Table 2b).[10] In the middle school sample, Math First respondents list even fewer names, on average, than Reading First respondents, though the difference is still not statistically significant. These results are particularly surprising, because Math First respondents answer the math name generator first, whereas Reading First respondents answer the math name generator only *after* answering the reading name generator.

### 3.1. Fatigue effects

The pattern of differences in out-degree between treatment groups are not consistent with fatigue effects. Recall that if fatigue effects are present in the survey, one would expect the number of names listed to decrease from the first name generator to the second. Contrary to expectation, such a decrease is observed in one treatment group, but not the other. If fatigue drives the differences in out-degree, one would not expect that the distribution of the total number of alters named in both generators to differ across treatment groups, because there is no reason for the two randomly assigned groups to differ in the amount of effort they are willing to exert. Such does not appear to be the case. In both samples we observe significant differences between the two treatment groups in the total number of alters named. Summarizing across samples, the Math First treatment group named 1.52 fewer alters, on average, than the Reading First treatment group (see Table 2c).

### 3.2. Satisficing effects

A satisficing respondent chooses how much effort to exert in responding to the second name generator by using the precedent of her response to the first name generator. If the true size of the second network is larger than the reported size of the first network, the respondent will list only as many names as she did in response to the first name generator, because such a response seems sufficiently complete. Effectively, the response to the first name generator creates a ceiling for the response to the second name generator.

The observed pattern of average differences in network size is consistent with satisficing behavior. Across samples and treatment groups, the average reading out-degree is larger than the average math out-degree. However, the magnitude of the difference is much smaller when the math name generator is posed first, which is consistent with the hypothesis that respondents are limiting the number of names they list in response to a later name generator based on the number of names they list in the initial generator.

To illustrate, look in detail at the elementary school sample. The Reading First treatment group begins by reporting an average 2.65 names in response to the first generator; for the second generator, they name an average of only 1.22 names, an average decrease of 1.43 names. Because the reading name generator tends to solicit more names than the math name generator, satisficing behavior does not create a constraint. In comparison, the Math First treat-

[10] Note that this difference is calculated by subtracting the math out-degree of the treatment group that answered the relevant name generator *second* from the treatment group that answered the name generator *first*.

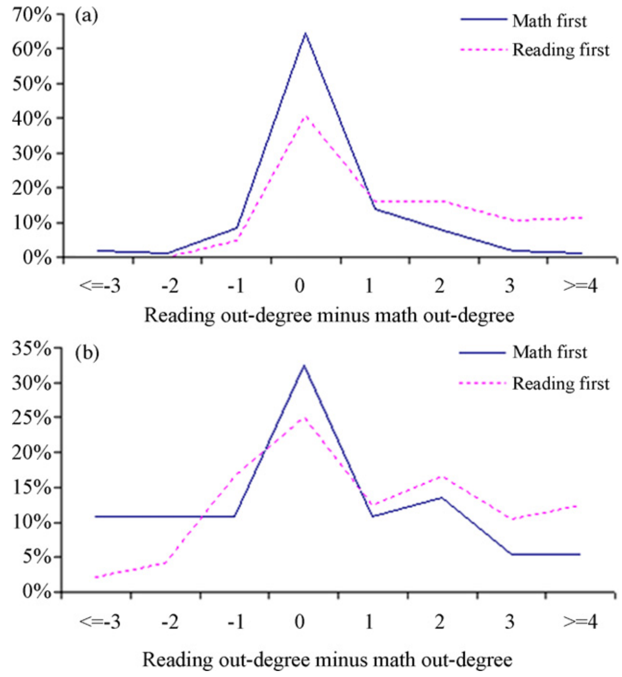

**Fig. 1.** (a) Elementary school sample. (b) Middle school sample.

ment group begins by listing an average of 1.13 names in response to the first name generator, similar to the number listed by the Reading First treatment group. One might then expect the Math First treatment group to list about 2.6 names in response to the second name generator, but this group lists an average of only 1.32 names in the reading name generator. This is consistent with satisficing behavior, where the small number of names listed in response to the first name generator leads respondents to list only a small number of names in response to the second name generator, creating a ceiling effect.

Satisficing effects are also evident in the differences between reading out-degree and math out-degree at the level of the individual respondent. Fig. 1a plots the frequency distribution of the difference between reading out-degree and math out-degree by treatment group for the Elementary school sample. Respondents in the Math First treatment group are likely to have a difference close to zero, because their response to the first name generator creates a ceiling for their response to the second name generator. Only 24% of respondents in the elementary school Math First treatment group listed more names in the reading generator than in the math generator, whereas in the elementary school Reading First treatment group, 54% of respondents listed more names in the reading generator than in the math generator. A similar pattern is observed in the middle school sample, though it is not as large in magnitude (see Fig. 1b).

### 3.3. *Non-redundancy effects*

The design of our survey does not allow robust tests for non-redundancy effects because we have no way of estimating the true level of multiplexity in the two networks that we measured—the extent to which respondents seek advice about both math and reading from the same alters. For purposes of description, we can only report on observed levels of overlap, in the form of Jaccard similarity coefficients. In the elementary school sample, 26% of all alters were named by the same respondent in both generators; in the middle school sample, 23% of alters were named by the same respondent in both generators.

### 3.4. *Cognitive priming*

The design of our survey does not allow tests for cognitive priming effects, because we do not have a method for determining how alters are organized in a respondent's memory.

### 3.5. *Question-scope redefinition*

Respondents understand the pragmatic meaning of questions by looking at the sequence of questions in a survey, as well other aspects of the design. In our survey, we observe how the scope of the second question could be redefined by the preceding name generator and name interpreter question. The name interpreter that follows the first name generator contains a bank of questions about the dimension of instruction for which the respondent seeks advice from each alter, questions asking the respondent to rate the frequency of their contact with each alter, and questions asking the respondent to rate the influence of the alter's advice on the respondent's practice. We examine two different types of question-scope redefinition, one that has to do with the design of the survey and one that has to do with respondents' understanding of the school subject areas that they teach.

First, the name interpreter questions provide additional, specific context that could influence the respondent's understanding of subsequent name generators. The respondent may recall the descriptions of different dimensions of instruction as examples of issues about which they have sought advice, almost as if the second name generator read: "In the past year, to whom have you gone for advice about teaching Mathematics, *for example*, *about deepening your content knowledge*, *planning or selecting course content and materials*, *approaches for teaching content to students*, *strategies specifically to assist low-performing students*, *or assessing students*' *understanding of the subject*?"

The five specific instructional dimension questions in the first name interpreter provide context that seems to be applied in answering the second name generator. In both treatment groups of both studies, the total number of the content areas checked increases from the first interpreter to the second interpreter, from an average of 2.76 to 2.90 (See Table 3).

While these differences are not large in magnitude, a suggestive trend appears in the Middle school sample. Across treatment groups, four out of five categories are checked with increased frequency in the second interpreter. Only the 6th "other" category is checked less frequently in the second interpreter. In the Middle school sample, the "other" category is checked 15% of the time in the first interpreter and 4% of the time in the second (see Table 3). The set of alters named in the second name generator appears to be a better fit for the categories of advice content, suggesting that the scope of the second question has been redefined by the content-area questions answered during the first name interpreter question. The set of alters named in the second name generator appears to be a better fit for the categories of advice content, suggesting that the scope of the second question may have been redefined by the content-area questions answered during the first name interpreter question. We note, however, that no clear pattern exists in the Elementary school sample.

The above analysis presents differences between the first and second sets of name interpreter data, averaging across treatment groups, but question-scope effects can also be analyzed by treatment groups. Differences between treatment groups in the number of content-areas checked are observed for the math network, though the pattern is unclear for the reading network. In response to the math network name interpreters, respondents in the Reading First treatment group checked 0.55 more content areas (out of five) per alter than did respondents in the Math First treatment group (Table 4b). Even though the Reading First treatment group

**Table 3**
Percent of alters with specific instructional dimension checked.

| Sample | Content-area | First interpreter | Second interpreter | Difference |
|---|---|---|---|---|
| Elementary schools | Deepening your content knowledge | 50% | 48% | −2% |
| | Planning or selecting course content and materials | 69% | 68% | −1% |
| | Approaches for teaching content to students | 64% | 70% | 5% |
| | Strategies specifically to assist low-performing students | 62% | 59% | −2% |
| | Assessing students' understanding of the subject | 51% | 57% | 5% |
| | Other | 11% | 4% | −7%[b] |
| | Total number of content-areas checked per alter[a] | 2.97 | 3.03 | 0.05 |
| Middle schools | Deepening your content knowledge | 44% | 44% | 0% |
| | Planning or selecting course content and materials | 58% | 65% | 7% |
| | Approaches for teaching content to students | 64% | 70% | 6% |
| | Strategies specifically to assist low-performing students | 43% | 48% | 5% |
| | Assessing students' understanding of the subject | 41% | 52% | 11%[b] |
| | Other | 15% | 4% | −11%[b] |
| | Total number of content-areas checked per alter[a] | 2.49 | 2.78 | 0.29[c] |
| Combined sample | Deepening your content knowledge | 47% | 46% | −1% |
| | Planning or selecting course content and materials | 64% | 67% | 2% |
| | Approaches for teaching content to students | 64% | 70% | 6% |
| | Strategies specifically to assist low-performing students | 54% | 54% | 0% |
| | Assessing students' understanding of the subject | 47% | 54% | 8%[b] |
| | Other | 13% | 4% | −9%[b] |
| | Total number of content-areas checked per alter[a] | 2.76 | 2.90 | 0.14 |

[a] Excludes the non-specific "other" category.
[b] Difference is significant at the 5% level according to a Fisher exact test.
[c] Distributions in first interpreter and second interpreter differ from one another at the 5% significance level according to a Mann–Whitney test.

**Table 4a**
Reading network: mean number of instructional dimensions checked per alter, by treatment group.

| Sample | Treatment group | | Difference (std. error) | Standardized *U* statistic[a] |
|---|---|---|---|---|
| | Reading First | Math First | | |
| Elementary schools | 3.11 | 2.69 | 0.42 (0.23) | 1.91 |
| Middle schools | 2.52 | 2.50 | 0.01 (0.28) | 0.16 |
| Combined sample[b] | | | 0.25 (0.18) | 1.05 |

[a] The Mann–Whitney *U* statistic tests the null hypothesis that the observed sample from each treatment group was drawn from a common distribution. Under the null hypothesis, the standardized Mann–Whitney *U* statistic is approximately normally(0,1) distributed.
[b] The estimates for the combined sample are calculated by taking the weighted average of the estimate from each sample, with weights inversely proportional to the variance of the estimated difference.

answered these math network name interpreters after filling in a previous set of name generators and interpreters, this group still checked more instructional dimensions, a pattern that is consistent with question-scope redefinition. In contrast, there is not a statistically precise difference in the number of alters checked in response to the reading network name interpreters. Respondents in the Reading First treatment group checked 0.25 more content areas (out of five) per alter than did respondents in the Math First treatment group, a difference that is not statistically significant (see Table 4a).

**Table 4b**
Math network: mean number of instructional dimensions checked per alter, by treatment group.

| Sample | Treatment group | | Difference | Standardized *U* statistic[a] |
|---|---|---|---|---|
| | Reading First | Math First | | |
| Elementary schools | 3.36 | 2.70 | 0.66 (0.26) | 2.49[b] |
| Middle schools | 2.86 | 2.42 | 0.45 (0.27) | 1.76 |
| Combined sample[c] | | | 0.55 (0.19) | 2.14[b] |

[a] The Mann–Whitney *U* statistic tests the null hypothesis that the observed sample from each treatment group was drawn from a common distribution. Under the null hypothesis, the standardized Mann–Whitney *U* statistic is approximately normally(0,1) distributed.
[b] The null hypothesis is rejected at the 5% level.
[c] The estimates for the combined sample are calculated by taking the weighted average of the estimate from each sample, with weights inversely proportional to the variance of the estimated difference.

## 4. Discussion and design considerations

In two samples collected using a multiple name generator survey that randomized the order of name generators, we strong find evidence of satisficing (rather than fatigue effects) and some evidence of question-scope redefinition. Evidence for non-redundancy effects and priming effects is not available in our experimental design. The effects for which we have found evidence are troubling because they are so closely related to the substantive questions that provoked our research. Specifically, satisficing effects and question-scope redefinition effects create biases that would cause us to reach opposite conclusions depending on the order in which the name generators and interpreters were posed.

In both samples, satisficing effects bias out-degree in a substantively meaningful way. If the purpose of our research were only to determine whether teachers sought more advice about reading or about math, we would reach opposite conclusions if we looked only at the Reading First treatment or only at the Math First treatment. Data from the Reading First treatment suggests that the average reading out-degree (and therefore network density) is significantly larger than the average math out-degree. On the other hand, data from the Math First treatment suggests that the difference between reading and math is much smaller, and statistically insignificant. If satisficing is creating bias in the number of names listed, as this evidence suggests, then one should look to the averages from the Reading First treatment group only, rather than from the entire sample, to find the best estimate of the true difference between reading out-degree and math out-degree.

Similarly, the question-scope effects we observe confound the possibility of comparing the different dimensions of instruction in the two subject-area networks. Suppose that we are interested in learning whether teachers seek advice about a broader array of dimensions of instruction in reading or in math. Because of the question-scope redefinition created by the sequence of the survey (generator, interpreter, generator, interpreter), we would have reached different conclusions depending on the order in which the subject areas were measured.

Based on earlier theory building and hypothesis generating work (Diamond and Spillane, 2006; Hayton and Spillane, 2007), we believe that the differences between treatment groups are driven by respondents' subject-specific thinking about advice-seeking. Question-scope redefinition effects would be observed if a respondent's understanding of the scope of advice that is sought for a particular subject is carried over to the respondent's interpretation of subsequent name generators. Our earlier work suggests that when elementary school staff interact about mathematics, conversations tend to focus on fewer dimensions of instruction compared to interactions about reading. If a respondent is asked about math first, she may apply this narrower understanding of math advice in responding to the reading name interpreter, and therefore check fewer dimensions of instruction. If she is asked about reading first, the broader understanding of reading advice is carried over to the math questions, so she checks more dimensions of instruction.

Our conclusions are limited in several of ways, some of which are suggestive of directions for further research. First, the design of our instrument does not allow us to test for non-redundancy effects or for cognitive priming effects. Both effects are strongly influenced by the amount of overlap between the networks being measured. Below, we suggest a survey design that would allow for tests of non-redundancy and priming.

Second, the analysis we have presented tests separately for the various types of question-order effects. In order to isolate the relative contribution of each effect (for example, to determine whether satisficing or cognitive priming is more important), an integrated model would be necessary.

Third, the validity and accuracy of any measurement depends not just on the instrument used to collect the data, but also on the particular statistic or metric that is applied to the raw data (Costenbader and Valente, 2003; Zemljic and Hlebec, 2005). We have focused primarily on the very basic measure of out-degree. Whether more complex metrics such as closeness, betweenness, or transitivity indices could be affected by question-order remains a subject of future work.

Fourth, our results may not generalize to multiple name generator surveys that measure different sets of criterion relationships. We have measured and attempted to compare two criterion relationships, advice about mathematics and advice about reading, that vary only in the school subject of interest. Both criterion relationships focus on the core work of school staff. Further methodological research is needed to examine question-order effects using criterion relationships that are less parallel. In the realm of organizational network analysis, sets of criterion relationships such as friendship, co-work, and advice-seeking should be tested. In the realm of personal network research, or studies of social capital, multiple name generators that measure instrumental support, emotional support, and social support should be tested.

Finally, we also urge caution in generalizing to other organizational settings. Research on network accuracy has suggested that the accuracy of name generator recall depends on the degree to which respondents have a well-developed structure for storing memories of other people (Freeman et al., 1987). Biases created by question-order effects may be lessened to the extent that name generators specify criterion relationships in social systems for which respondents have good mental models. For example, corporate headhunters or community organizers might very likely have good mental models for their contacts, since they make constant use of them. Teachers' mental models of advice sources may be developed and accessed very differently than, for example, teenagers' mental models of social support. Further research on cognitive models used in name generator recall should therefore be domain-specific, attending to the relationship between the research setting and instrument design.

Based on our findings, we conclude with some general suggestions about instrument design for capturing multiplex network data, addressing the relative merits of rosters versus name generators, considerations about the relationship between criterion relationships, and the sequencing of name interpreter questions.

For measuring single networks, others have recommended using complete roster methods whenever possible (Brewer, 2000). For multiple criterion relationships, roster methods would also seem to have an advantage; by not asking respondents to recall names from memory, attendant problems of fatigue, satisficing, non-redundancy, and priming effects could be avoided. However, roster-based methods may involve a considerable response burden, which must be weighed against the advantages of the design. In choosing roster-based methods over name generator surveys, the researcher might also be trading in one set of context effects for another. When posing a set of questions about each alter in an organization, response effects such as fatigue or satisficing might come into play based on the order of the alters in the roster. Further methodological research is needed to test the validity of roster-based designs for multiple name generator surveys.

In situations where roster-based methods are not feasible, multiple name-generator surveys should be designed with careful attention to the relationship between the criterion relationships of interest. The researcher should consider the relative density of the networks likely to be generated, the likely degree to which criterion relationships will overlap with one another, and how the criterion relationships may be perceived in relation to one another.

In studies where relative network size is of primary interest, minimizing the possibility of fatigue or satisficing effects is a key concern. If the mode of data collection permits, one might consider randomizing the order in which the name generators are presented, so that the extent of fatigue or satisficing effects can be quantified. Such split-sample experiments can be a highly useful and revealing form of survey pre-testing (Fowler, 2004).

In studies where the multiplexity of several criterion relationships is of primary interest, confounding processes such as non-redundancy effects or priming effects should be controlled. Interpreting the results of a multiple name generator survey, one might easily assume that the non-inclusion of a given alter in response to a name generator means that the alter does not fit the specified criterion relationship. However, if question-order effects are likely to bias the process of recalling names from memory, one should be wary of this assumption. Instead, the task of generating names should be separated from the task of interpreting information regarding those names. Alternately, the survey designer may be able to temper non-redundancy effects by including specific instructions to survey respondents, such as "If applicable to this question, please also include the names of people that you have listed in response to previous questions."[11]

To control for the possibility of non-redundancy or priming effects, name generators should all be run first, using specific criterion relationships or more general ones, and prompting the respondent to keep searching her memory if appropriate. Once a set of alter names has been generated, name interpreter questions could be posed that ask the respondent to classify the alter into one or more of the criterion relationships of interest. A similar approach has been applied in surveys that collect ego-centric net-

[11] We thank an anonymous reviewer for suggesting this approach.